# Practical Estimation of Diversity from Abundance Data

Eric Marcon

# Practical Estimation of Diversity from Abundance Data

Eric Marcon[1]*

**Abstract**
Measuring biodiversity requires empirical techniques to effectively estimate it from real data. The well-known underestimation of the number of species applies to low orders of diversity in general. I test nine estimators including three new ones on geometric and lognormal distributions that represent realistic, hyper-diverse communities. The best two estimators allow a good estimation of diversity of orders over 0.5, even when the sampling effort is low. I provide criteria to choose the estimator and the necessary code in the R package entropart.

**Keywords**
Biodiversity, HCDT entropy, Phylodiversity

[1]*AgroParisTech, UMR EcoFoG, CNRS, Cirad, INRA, Université des Antilles, Université de Guyane, Campus agronomique, BP 316, F-97310 Kourou, French Guiana.*
***Corresponding author**: Eric.Marcon@ecofog.gf

## Contents



## Introduction

Measuring biodiversity requires both a robust theoretical framework (Patil and Taillie, 1982) and empirical techniques to effectively estimate the theoretical variables with real data (Beck and Schwanghart, 2010). In this paper I focus on species-neutral measures of diversity based on HCDT entropy (Havrda and Charvát, 1967; Daróczy, 1970; Tsallis, 1988) that fulfill the first requirement. Entropy measures the average surprise brought by observing individuals of a community. Surprise is a decreasing function of probability dropping to 0 when probability is 1. HCDT entropy uses a parameterized surprise function that is the deformed logarithm of order $q$ of the reciprocal of probability(Marcon *et al.*, 2014a). Traditional measures of diversity, namely the number of species as well as Shannon's and Simpson's indices, are special cases of the HCDT entropy for values of $q$ equal to 0, 1 and 2. HCDT entropy should be transformed into Hill numbers (Hill, 1973) for better interpretation of the value of diversity as an effective number of species (Jost, 2006). Hill numbers are simply the deformed exponential of HCDT entropy (Marcon *et al.*, 2014a). Rather than focusing on a single value of $q$, a profile of diversity, *i.e.* a plot of diversity against $q$, can be built (Tothmeresz, 1995). Low values of $q$ (starting from 0) give much importance to rare species, whilst higher values (usually up to 2) focus on abundant species. Negative values of $q$ are not used because of poor mathematical properties of their entropy (Beck, 2009), and values over 2 generally bring little more information. Ordering communities in terms of diversity requires that their profile do not cross (Tothmeresz, 1995); else, declaring a community more diverse than another only holds for a range of values of $q$ reflecting the importance given to rare or frequent species (Lande *et al.*, 2000).

To plot those profiles, diversity must be estimated from the data. Estimation bias (I follow the terminology of Dauby and Hardy, 2012) is a well-known issue (Marcon *et al.*, 2014a). Real data are almost always samples of larger communities, so some species may have been missed. The induced bias on the Simpson entropy is

smaller than on the Shannon entropy because the former assigns lower weights to rare species, *i.e.* the sampling bias is even more important when $q$ decreases. Another estimation bias has been widely studied by physicists who generally consider that all species of a given community are known and their probabilities quantified. Their main issue is not at all missing species but the non-linearity of entropy measures (see Bonachela *et al.*, 2008, for a short review). Estimating probabilities at power $q > 0$ by the power of their estimator is an important source of underestimation of entropy. The need for corrections has generated a considerable literature in ecological statistics and statistical physics.

In this paper, I test the performance of the state-of-the-art estimators when applied to the kind of data ecologists have to deal with. I start with simulated distributions that have the advantage of being easily manipulated to generate various sampling intensities and evaluate the bias and root mean square error (RMSE) of the estimators. I address the classical models of the literature, namely the lognormal and the geometric distributions. The lognormal distribution describes, at least roughly, many hyperdiverse ecosystems even though the link between its statistical success and the underlying ecological mechanisms is poorly documented (Tokeshi, 1993). The geometric distribution is a far more difficult case because it is very uneven: the frequency of rare species is several orders of magnitude smaller than that of the frequent ones, making it impossible to observe with reasonable sampling effort (Haegeman *et al.*, 2013). I apply the best-known and best-performing estimators, including three new ones, to those distributions and two actual forest data sets. My purpose is to provide recommendations about the estimation technique to chose when facing different types of data and draw general conclusions about the possible accuracy of diversity estimation.

Phyloentropy is the sum of HCDT entropy along an ultrametric tree (Marcon and Hérault, 2015a) so estimating it reduces to estimating HCDT entropy. Phylodiversity is the deformed exponential of phyloentropy. In short, estimating phylodiversity relies on the methods presented here so I will focus on species-neutral diversity for clarity.

I used the package *entropart* (Marcon and Hérault, 2015b) for R (R Development Core Team, 2015) for all tests. The R code necessary to reproduce all results is in the electronic appendix.

## 1. Methods

Consider a community of species indexed by $s = 1, 2, \ldots, S$. $n_s$ is the number of individuals of species $s$ sampled in the community, $n = \sum_s n_s$ the total number of sampled individuals. The (unknown) probability $p_s$ for an individual to belong to species $s$ is estimated by $\hat{p}_s = n_s/n$.

The number of species represented by $\nu$ individuals in the sample of size $n$ is $s^n_\nu$, so $s^n_0$ if the (unknown) number of unobserved species considering the sampling effort. $s^n_\nu$ is considered as a realization of the random variable $S^n_\nu$ so it is used to estimate its expectation $\mathbb{E}(S^n_\nu)$.

$\pi_\nu$ is the sum of the probabilities $p_s$ of species represented by $\nu$ individuals.

The deformed logarithm formalism (Tsallis, 1994) is very convenient to manipulate entropies. The deformed logarithm of order $q$ is defined as:

$$\ln_q x = \frac{x^{1-q} - 1}{1-q} \tag{1}$$

It converges to the natural logarithm when $q \to 1$.

The inverse function of $\ln_q x$ is the deformed exponential:

$$e^x_q = [1 + (1-q)x]^{\frac{1}{1-q}} \tag{2}$$

### 1.1 Sample coverage

The sample coverage (Good, 1953) is the probability for an individual in the community to belong to a species observed in the sample. It equals the sums of the probabilities of the observed species. It is an essential tool for diversity estimation because it is included in some estimators (*e.g.* Chao and Shen, 2003) and it allows the evaluation of the completeness of sampling (Chao and Jost, 2012). Its estimator given by Good is:

$$\hat{C} = 1 - \frac{s^n_1}{n} \tag{3}$$

It is biased (Zhang and Huang, 2007), because:

$$C = 1 - \frac{\mathbb{E}(S^n_1) - \pi_1}{n} \tag{4}$$

Good's estimator neglects the term $\pi_1$, the sum of the probabilities of singletons. It was built from Turing's frequency formula relating the average probability of species observed $\nu$ times to the number of species observed $\nu+1$ and $\nu$ times. This formula has been improved by Chao *et al.* (Chao and Shen, 2010; Chiu *et al.*, 2014) to estimate $\pi_1$. Estimating the number of species by the Chao1 estimator (Chao, 1984), Chao and Shen (2010) obtained an improved estimator of the sample coverage:

$$\hat{C} = 1 - \frac{s^n_1}{n}\left[\frac{(n-1)s^n_1}{(n-1)s^n_1 + 2s^n_2}\right] \tag{5}$$

This estimator has been further used by Chao and Jost (2015) to derive an estimator of entropy (see below).

An almost unbiased estimator has been derived using the information provided by the whole distribution (Chao *et al.*, 1988; Zhang and Huang, 2007):

$$\hat{C} = 1 - \sum_{\nu=1}^{n} (-1)^{\nu+1} \binom{n}{\nu}^{-1} s^n_\nu \tag{6}$$

I use it in this paper.

### 1.2 Estimators of entropy

The existing estimators and the new ones proposed here can be classified into four main methods. The simplest one just consists of plugging the estimator of $p_s$, *i.e.* $\hat{p}_s = n_s/n$, into the definition of diversity to evaluate to obtain a so-called plug-in estimator, sometimes named naive estimator. The plug-in estimator of HCDT entropy of order $q$ is:

$$ {}^q\hat{H} = \sum_s \hat{p}_s \ln_q \frac{1}{\hat{p}_s} \tag{7} $$

The plug-in estimator is useless in hyper-diverse communities because it severely underestimates diversity because of unobserved species and of the non-linearity of estimators.

Recent progress have been made in estimating the actual distribution of the probability of species by fitting a model of their distribution to the data. The distribution of the unobserved species can be added if their number is estimated and a distribution form is chosen. Chao *et al.* (2015) used a two-parameter model based on the estimation of the generalized sample coverage (not detailed here), estimated the total richness with the Chao1 estimator and modeled the unobserved species as a geometric distribution to unveil the complete rank-abundance distribution of an observed community. They applied the plug-in estimator this distribution: I'll call it the "Chao-unveiled" estimator.

The Chao1 estimator has been built according to the same theoretical approach as that of the unveiled rank-abundance distribution. It is a lower-bound estimator of the number of species. It has been improved by Chiu *et al.* (2014) who slightly reduced its negative bias with the iChao1 estimator, integrating species represented by 3 and 4 individuals. I define the "iChao-unveiled" estimator as a variation on the "Chao-unveiled" estimator, where richness is estimated by the iChao1 estimator.

The jackknife estimator (Burnham and Overton, 1979) has shown good performances to estimate richness when the sampling effort is too low for the Chao1 estimator to perform well (Brose *et al.*, 2003) even though it actually lacks theoretical support Cormack (1989). I test the use of the jackknife estimator, whose order is selected according to the data, to define the "jackknife-unveiled" estimator. Using the jackknife estimator to unveil the tail of the abundance distribution was not the intention of Chao *et al.* (2015) because it is not consistent with their theoretical framework. It must be seen here as a merely empirical tool.

The second method relies on the Horvitz and Thompson (1952) estimator of the weighted sum of a function of its elements, say $\sum_s p_s f(s)$ when some of them are not observed. An unbiased estimator of the sum is obtained when each term is divided by its probability to be observed $1-(1-p_s)^n$. Chao and Shen (2003) proposed to combine it with the estimator of the sample coverage: conditionally to the set of observed species, an unbiased estimator (Ashbridge and Goudie, 2000) of $p_s$ is $\tilde{p}_s = \hat{C}\hat{p}_s$. Chao and Shen estimated the Shannon entropy; the method has then been extended to HCDT entropy (Marcon *et al.*, 2014a) and similarity-based diversity (Marcon *et al.*, 2014b):

$$ {}^q\tilde{H} = \sum_{s=1} \frac{\hat{C}\hat{p}_s \ln_q \frac{1}{\hat{C}\hat{p}_s}}{1-\left(1-\hat{C}\hat{p}_s\right)^n} \tag{8} $$

A further progress can be done by replacing the conditional estimator of probabilities $\tilde{p}_s = \hat{C}\hat{p}_s$ by that of Chao *et al.* (2015). Since the improved probability estimator depends on the generalized sample coverage, I'll call the improved Chao-Shen estimator the "generalized coverage" estimator.

The third method has been derived by Grassberger (1988) who gave a reduced-bias estimator of the value of an integer at power $q$. $p_s^q$ is written as $n_s^q/n^q$ and $n_s^q$ is estimated (Marcon *et al.*, 2014a) as:

$$ \widetilde{n_s^q} = \frac{\Gamma(n_s+1)}{\Gamma(n_s-q+1)} + \frac{(-1)^n \Gamma(1+q)\sin\pi q}{\pi(n+1)} \tag{9} $$

The estimator of $p_s^q$ is simply $\widetilde{p_s^q} = \widetilde{n_s^q}/n^q$. It is plugged into the formula of entropy to obtain the Grassberger estimator:

$$ {}^q\tilde{H} = \frac{1-\sum_s \widetilde{p_s^q}}{q-1} \tag{10} $$

The last method has been the subject of an important literature in the last ten years. A review can be found in Chao *et al.* (2013), Appendix A. It relies on the estimation of $h_q = \sum_s p_s^q$. $h_q$ can be written as the following sum:

$$ h_q = \sum_{r=0}^{\infty} \binom{q-1}{r} (-1)^r \zeta_r \tag{11} $$

$\zeta_r$ is the generalized Simpson entropy $\sum_s p_s(1-p_s)^r$ defined by Zhang and Zhou (2010). The first $n$ elements of the sum, denoted $\tilde{h}_q$, can be estimated with no bias (Zhang and Grabchak, 2014):

$$ \tilde{h}_q = \sum_{s=1}^{S} \hat{p}_s \sum_{v=1}^{n-n_s} \left[ \prod_{i=1}^{v} \frac{i-q}{i} \prod_{j=1}^{v} \left(1-\frac{n_s-1}{n-j}\right) \right] \tag{12} $$

Zhang (2013) shows that the bias due to ignoring the remaining terms is asymptotically normal and decays exponentially fast. I'll call the Zhang and Grabchak (2014) estimator the one based on $\tilde{h}_q$:

$$ {}^q\tilde{H} = \frac{1-\tilde{h}_q}{q-1} \tag{13} $$

Some attempts have been made to estimate the remaining bias (Zhang and Grabchak, 2013). The most achieved one is that of Chao and Jost (2015), completing Chao *et al.* (2013). It relies on two assumptions: the total number of species is estimated by the Chao1 estimator and the actual probabilities of unobserved species can be estimated all equal. A consequence is that the estimator of the average probability of species sampled once also equals the probability estimator of unobserved species. Its value is noted $A$. It is $2s_2^n/[(n-1)s_1^n+2s_2^n]$ if singletons and doubletons are present or $2/[(n-1)(s_1^n-1)+2]$ if doubletons are missing. The Chao-Wang-Jost estimator of HCDT entropy is:

$$ \begin{aligned} {}^q\tilde{H} = \frac{1}{q-1}[1-\tilde{h}_q \\ -\frac{s_1^n}{n}(1-A)^{1-n}\left(A^{q-1}-\sum_{r=0}^{n-1}\binom{q-1}{r}(A-1)^r\right)] \end{aligned} \tag{14} $$

In absence of singletons and doubletons, $A$ is set to 1 and the estimator is identical to that of Zhang and Grabchak.

### 1.3 Confidence intervals

Two methods allow the evaluation of confidence intervals: asymptotic, closed forms are available for some estimators, or bootstrapping is required in the general case.

Esty (1983), completed by Zhang and Huang (2007), showed that the estimator of sample coverage (eq. 6) is asymptotically normal with the following confidence interval:

$$ C = \hat{C} \pm t_{1-\alpha/2}^n \frac{\sqrt{s_1^n\left(1-\frac{s_1^n}{n}\right)+2s_2^n}}{n} \tag{15} $$

Where $t_{1-\alpha/2}^n$ is the quantile of a Student distribution with $n$ degrees of freedom at the risk threshold $\alpha$, here 1.96 for all sample sizes and $\alpha = 5\%$.

The Zhang-Grabchak estimator is also asymptotically normal and comes with an asymptotic confidence interval (Zhang and Grabchak, 2014) implemented in the package *EntropyEstimation* (Cao and Grabchak, 2014).

The theoretical distribution of other estimators is unknown. They must be built by bootstrap techniques: the observed community is re-sampled, say 1000 times, and entropy is calculated each time. The $\alpha/2$ and $1-\alpha/2$ quantiles of the distribution of entropy are the bounds of the confidence interval. The issue of re-sampling a community is the same as that of sampling it: rare species are often eliminated, so the entropy is underestimated. Starting from the whole community, a first estimation bias is caused by sampling it. The estimators presented here aim at correcting it. When this observed community is re-sampled, a second estimation bias appears. Estimating the entropy of re-sampled communities with bias correction yields, on average, the entropy of the observed community estimated by the plug-in estimator (Marcon *et al.*, 2012): if the estimator works well, it eliminates the second estimation bias but it cannot address the first one. The solution to this problem is simply to recenter the entropy distribution of re-sampled communities around the value of the entropy of the observed community (Marcon *et al.*, 2012; Chao and Jost, 2015).

The re-sampling technique may just consist of drawing individuals in the observed community with replacement, or, equivalently, drawing a community in a multinomial distribution respecting the size and probability distribution of the observed community (Marcon *et al.*, 2014a). A more sophisticated technique has been proposed by Chao and Jost (2015). Given the sample size, the probability distribution of observed species can be estimated more accurately than by the estimator $\tilde{p}_s = \hat{C}\hat{p}_s$ which underestimates the probability of rare species (Chao *et al.*, 2015). A better estimate of the probabilities is used (actually, a simplified version of that of the unveiled estimators above) and completed by an estimation of the number of unobserved species, whose probabilities are assumed identical. Despite these extra efforts, the distribution of the entropy of re-sampled community still has to be recentered.

### 1.4 From entropy to diversity

All entropy estimations are finally transformed into diversity values to be interpretable (Jost, 2006). It is not correct to recenter the confidence interval of diversity estimations because of the non-linearity of the transformation of entropy into diversity (Marcon *et al.*, 2012). The correct process consists of evaluating entropy with its confidence interval and make the final exponential transformation of all values into diversity.

### 1.5 Typical distributions

Comparing the performance of estimators requires simulations of realistic communities. I chose to focus on two opposed models making sense in ecology. The lognormal distribution (Preston, 1948) fits well species-rich communities for several reasons, including populations dynamics (Engen and Lande, 1996), niche apportionment (Bulmer, 1974), or even statistical physics arguments (Pueyo *et al.*, 2007; Dewar and Porté, 2008). It is often well fitted

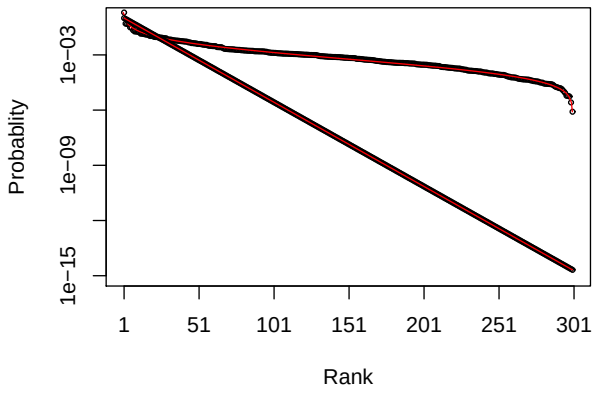



**Figure 1.** Rank-Abundance curves of 300 species following a lognormal (top curve) or a geometric distribution (straight line). The red lines are the fitted models.

empirically (Tokeshi, 1990) even though it has been questioned theoretically (Williamson and Gaston, 2005). The local community distribution according to the neutral theory (Volkov *et al.*, 2003) is not lognormal but departs from it very moderately. The logarithm of the species probabilities follows a Gaussian distribution.

The geometric series model (Motomura, 1932; Whittaker, 1972) generates far more uneven species distributions. In this model, the first species is represented by a part $p$ of the total resources. The second one has the same part $p$ of the remaining resources, and so on. Finally, probabilities are normalized to be proportional to the resources taken.

I generated four artificial communities following those distributions. Figure 1 presents a lognormal one, with log-standard-deviation equal to 2 (typical of the distribution of tree species in a rainforest) and a geometric distribution with parameter $p = 0.1$. Both contain 300 species. The other two distributions have identical parameters except for the number of species augmented to 600.

### 1.6 Evaluation of the performance of estimators

The performance of each estimator was calculated as its average relative bias on all values of $q$ (*i.e.* the average difference between the mean simulated entropy and its true value) and it Root Mean Square Error (RMSE, *i.e.* the square root of the sum of the squared bias and the variance, divided by the true value). The true entropy of each reference distribution was calculated with the known values of $p_s$. For each reference distribution, 1000 random samples of the chosen size were drawn in a multinomial distribution respecting the reference probabilities $p_s$. Entropy was calculated for $q$ between 0 and 2. The average entropy and its first and last 2.5% quantiles were retained to build the profile and its confidence envelope (which is quite different from that of the estimation of real communities). Finally, entropy was transformed into diversity to be plotted.

## 2. Results

I drew multinomial samples of various sizes in the chosen species distributions, simulating a real, independent sampling of individuals. Sample sizes are between 200 and 5000 individuals to cover a range from obvious undersampling to a high-effort inventory: 5000 individuals correspond to 9 to 10 ha of forest.

### 2.1 Sample coverage

I first evaluated the performance of the estimator of sample coverage. 2 communities of each size between 200 and 5000 individuals were sampled in each typical distribution. The real and estimated sample coverages are compared on figure 2. The estimation of sample coverage is very efficient. A model II linear regression (Legendre, 2014) validated the accuracy of the estimation.

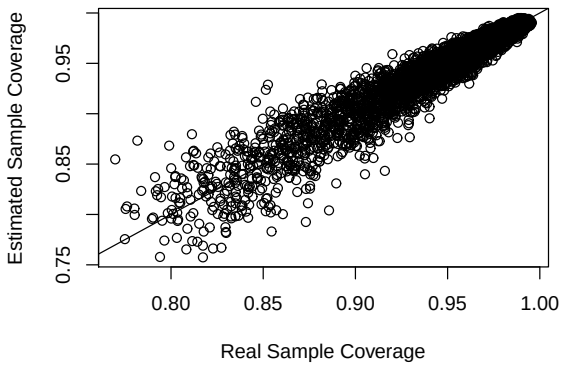
Conditionally to the sample size, the relation vanishes but the average estimation is very close to the average actual value: as predicted by the theory, the estimation bias is very small (Figures 7 and 8).

### 2.2 Entropy and diversity

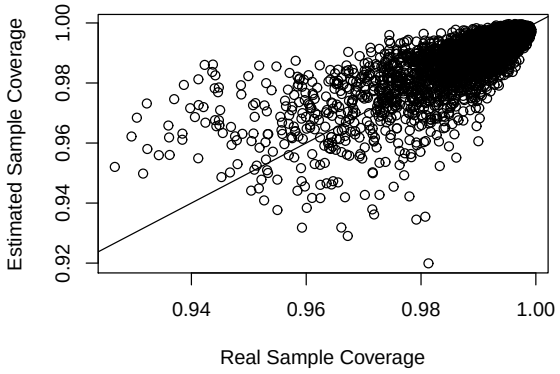
I estimated the entropy profiles of the lognormal and geometric distributions of 300 and 600 species, sampled at 4 different intensities (200, 500, 1000 and 5000 individuals), by 9 estimators: Chao-Shen, Grassberger, Chao-Wang-Jost, Zhang-Grabchak, Generalized Coverage, the three unveiled and the Plug-in. The diversity profiles are plotted in the electronic appendices.

The root mean square error of the estimators is shown on figure 3 for the lognormal and the geometric distributions with 300 species when 1000 individuals are sampled, a typical tropical forest inventory of trees.

Unsurprisingly, the plug-in estimator is severely biased and has the poorest results in the tests. The Chao-Wang-Jost estimator systematically outperforms the Zhang-Grabchak estimator (which actually performs little better than the plugin-estimator here) by construction. Its complementary estimation is not paid by increased variance. The Grassberger estimator is totally inefficient for low values of $q$ as already noticed by Marcon *et al.* (2014a). The generalized coverage estimator outperforms Chao-Shen because of its better estimation of conditional probabilities. The Chao-unveiled estimator is almost confused with the Chao-Wang-Jost estimator. Both are outperformed by the iChao-unveiled estimator because it improves the estimation of the number of

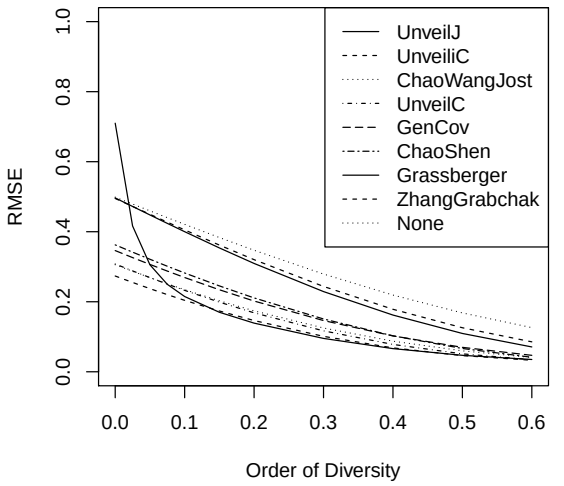



**(a)** Lognormal distribution. The estimated sample coverage is 0.991 times the real one plus 0.009, with an $R^2$ value around 94%.

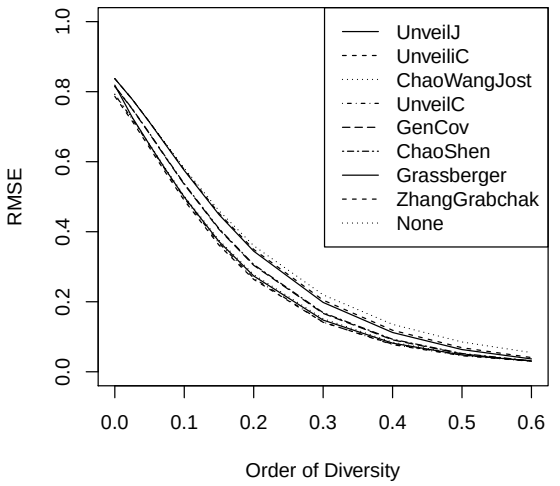



**(b)** Geometric distribution. The estimated sample coverage is 1.046 times the real one minus 0.046, with an $R^2$ value around 69%

**Figure 2.** Estimated *vs* real sample coverage of simulated samples of a (a) lognormal or (b) geometric distribution of 300 species. Sample sizes are between 200 and 5000 individuals. The line represents the fit of a model II (Major Axis method) linear regression.




**(a)** Lognormal distribution.




**(b)** Geometric distribution

**Figure 3.** Estimated relative RMSE of the estimators of diversity based on 1000 samples of 1000 individuals of each typical distribution (lognormal and geometric) of 300 species. The RMSE is normalized by the actual diversity. It is quite high for low orders of diversity, especially for the geometric distribution. Values of $q$ over 0.6 are not shown because all estimators perform similarly well. The legend lists the estimators in the increasing order of RMSE for $q > 0.1$, where the estimator with the lowest RMSE is the jackknife-unveiled one, closely followed by the iChao-unveiled and Chao-Wang-Jost. Close to $q = 0$, the jackknife-unveiled estimator has a higher variance making it the least reliable estimator.

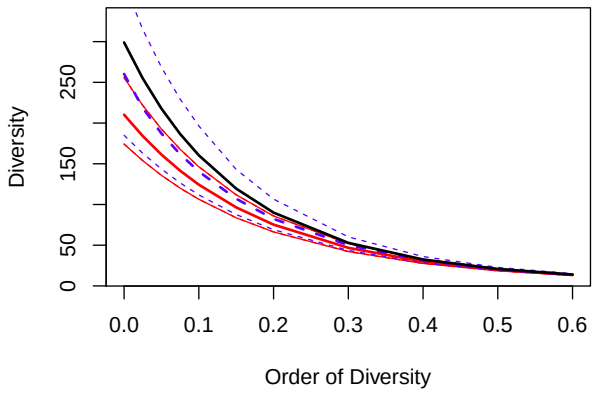



**Figure 4.** Diversity profiles estimated from 1000 random samples of 1000 individuals from a lognormal community of 300 species. The bold, black line represents the real diversity (starting from ${}^0D = 269$). The jackknife-unveiled estimator is plotted by the blue, dashed line (${}^0\hat{D} \approx 250$). Its confidence interval (blue dots) is very wide. The Chao-Wang-Jost estimator (red, bold line: ${}^0\hat{D} = 209$) is more biased downard but its confidence interval (red, solid lines) is much smaller.

species. The jackknife-unveiled estimator is more flexible than the previous ones to estimate the number of species. The order of the jackknife estimator it uses changes between simulations, causing an excessive variance for $q < 0.1$. It performs best for higher orders of diversity.

Results are consistent whatever the model. The general pattern is a poor estimation of low orders of diversity, and a quite accurate estimation of high orders, as previously shown by Haegeman *et al.* (2013). The RMSE varies a lot according to the model.

I am continuing the analysis with the best two estimators: Chao-Wang-Jost and jackknife-unveiled, ignoring the iChao-unveiled estimator which takes place between them but is too similar to the jackknife-unveiled to bring decisive arguments for the discussion. Figure 4 shows their profiles for a 1000-individual sample of a lognormal distribution of 300 species, with their confidence intervals.

## 3. Discussion

The underlying distribution of species is the most important determinant of the success of diversity estimation: the estimation bias of heavy-tailed distributions decays more slowly when the sample size is increased (Zhang and Grabchak, 2013). Estimating the low-order diversity of a sample from a geometric distribution is all but impossible (Haegeman *et al.*, 2013) but the low-order diversity of lognormal communities can be estimated meaningfully when the sample size is sufficient. Empirically, it is not possible to discriminate a severely-censused geometric distribution and a lognormal one (Tokeshi, 1993): both models fit well since most of the difference is contained by the unobserved tails of the distributions. So, theoretical, ecological arguments about the actual distribution of the community are necessary to decide whether an estimation of diversity is reliable.

Diversity of order over 0.5 is pretty well estimated in the context of this paper. Haegeman *et al.* showed that this remains true for $q \geq 1$, even when geometric communities of millions of species with parameter 0.5 (the most abundant species takes half the resources, the second one a quarter and so on) are addressed.

### 3.1 The sample coverage is not always the good indicator of the quality of estimation

The sample coverage can not be used as a proxy for how much an estimate of diversity can be relied upon. At the same sampling effort, the sample coverage appears to be higher for the geometric distribution (Figures 7 and 8). Far more species are not sampled than in a lognormal distribution, but their total probability is smaller. For example, samples of 200 individuals drawn in 300-species geometric and lognormal communities yield an average estimation of 54 and 149 species by the jackknife-unveiled estimator, but the respective sample coverages are over 95% for the geometric distribution versus around 81% for the lognormal one.

The estimation bias is thus much greater for low orders of diversity even though the sample coverage is higher. Chao and Jost (2012) argue in favor of the sample coverage as a better measure of the sampling effort than the sample size. I agree as long as the underlying distribution of communities is the same: then, standardizing the sampling effort by the sample coverage is pertinent.

### 3.2 Comparing the diversity of real communities with different distributions remains untractable

When the number of species of the theoretical distributions is doubled, everything else equal, the sampling bias increases (compare figures 14 and 18). With the same sampling effort, the coverage of the lognormal distribution decreases (compare figures 7 and 8, left columns). Doubling the effort brings both the sample coverage and the bias back to their previous level, with a reduced variance (compare figures 18c and 14e).

This is a very simple and intuitive behavior, but it is completely different with the geometric distribution: the sample coverage does not change when richness is doubled (compare figures 7 and 8, right columns) because the probabilities of the 300 rarest species are negligible. Doubling the sample size does not restore the bias level (compare figures 18d and 14f). An extensive and

rigorous analysis of the influence of the parameters of the theoretical distributions (beyond manipulating the number of species) is not the scope of this paper, but this simple example shows that no general and simple rules are available to compare the low-order diversity of communities of different nature.

### 3.3 Estimating the number of species is the critical step

The lower $q$, the more difficult the estimation is, but the estimation of the number of species has been long studied and simple rules of decision have been proposed (Burnham and Overton, 1979; Brose *et al.*, 2003) to chose the most appropriate order of the jackknife estimator. Burnham and Overton derived a selection procedure to obtain the order allowing to minimize the RMSE of the estimation of the number of species. It is implemented in the package *SPECIES* (Wang, 2011) for R. Brose *et al.* showed (empirically) hat the first-order jackknife is selected when the sample completeness (terminology by Beck and Schwanghart, 2010), *i.e.* the proportion of observed species $(S-s_0^n)/S$ is over 3/4 (precisely 74% in their paper). When it is less, higher orders have less bias but more variance. It is easy to estimate the number of species of an actual sample this way and compare it to the Chao1 estimator. If both coincide, the Chao-Wang-Jost estimator will perform well for the whole profile: its value at $q=0$ is that of Chao1. Else, the jackknife-unveiled estimator will be the best choice since its value at $q=0$ is the optimal-order jackknife. If one does not want to rely on the jackknife estimator for some reason, such as its poor theoretical support, the iChao-unveiled estimator is a reasonable compromise as a lower bound estimation.

### 3.4 Better, but probably not much better, estimators may be derived

The most promising ways of research according to the present results are a better estimation of the remaining bias of the Zhang-Grabchak estimator and the improvement of the distribution modeling of the unveiled estimators. The first approach is that of the Chao-Wang-Jost estimator, which is limited by its estimation of the number of species (the lower bound, Chao1 estimator). The price for releasing this constraint is losing the elegant, closed form of the estimator allowed by appropriate approximations of the infinite sum of the unknown elements of eq. (11) for a numeric approximation.

The distribution of species is modeled with two parameters in the unveiled estimators. This can be refined by extending the technique presented by Chao *et al.* (2015) to higher orders of sample coverage. In both cases, better fitting the data to reduce the bias has its limits because the variance of estimation is likely to increase (Bonachela *et al.*, 2008). So, the estimators presented here may not be far from the optimum trade-off (less bias with the jackknife-unveiled estimator, less variance with Chao-Wang-Jost).

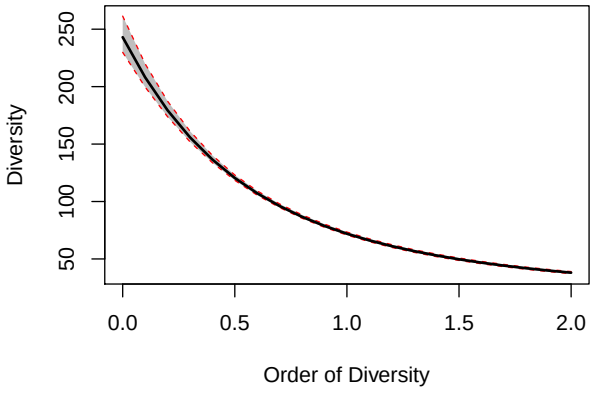



**Figure 5.** Estimated diversity profile of the tree species of the BCI 50-ha plot. The shaded zone is the 95% confidence interval of the estimation.

## 4. Application to real data

I now estimate the diversity of two real forest plots. The first case is Barro Colorado Island's 50-ha plot of tropical forest (Hubbell *et al.*, 2005), whose inventory data of trees over 10 cm diameter at breast height are available in the package *vegan* (Oksanen *et al.*, 2012) for R. 225 species have been sampled, with a quite good fit to a lognormal distribution. The sample size is over 20000 individuals, the sample coverage is over 99.9%. Estimating the number of species with the Chao1 (239 species) or the jackknife 1 (244) estimators gives very similar results. This is an unusually large dataset, whose diversity estimation (Figure 5) is quite easy.

The best estimator is Chao-Wang-Jost since the Chao1 estimator is appropriate for the number of species. The 95% confidence interval of the estimation is built by resampling according to the technique by Chao and Jost (2015). It is very small due to the abundance of data.

The second example takes place at the other extreme of sampling intensity. A 1-ha plot (plot 18) of tropical forest in the experimental forest of Paracou (Gourlet-Fleury *et al.*, 2004), French Guiana, has been inventoried. Data are available in the package *entropart* for R. Only 481 trees over 10 cm diameter at breast height have been sampled. They belong to 149 species. The sample coverage is 84.6±4.4%. The estimated number of species is 254 according to Chao1, but the appropriate Jackknife estimator (of order 3) returns 309 species. Clearly, the sampling effort is not sufficient for an accurate estimation: the sample coverage is too low and the estimation of the

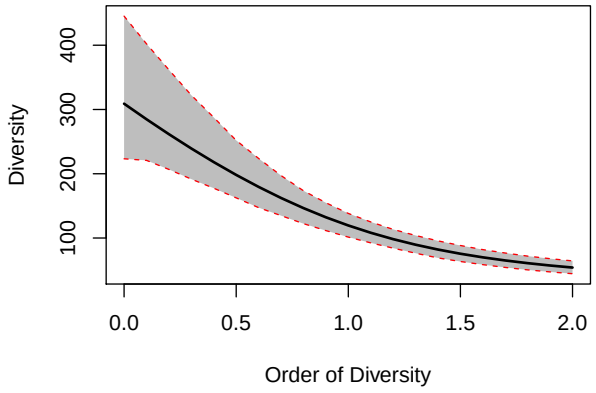



**Figure 6.** Estimated diversity profiles of the tree species of the Paracou 1-ha plot #18. The shaded zone is the 95% confidence interval of the estimation.

number of species too uncertain. With no doubt, the Chao-Wang-Jost estimator will severely underestimate diversity.

The jackknife-unveiled estimator is the best choice. Its confidence interval is very wide up to $q = 0.3$. Over $q = 0.5$, the simulations above showed that the estimator has a very low variability, so the confidence interval is due to the uncertainty of the sampling only. At lower orders of diversity, the estimator's uncertainty amplifies it so the estimation is not reliable. In this case, the very little accuracy of the jackknife-unveiled estimator (the number of species is estimated between 237 and 439) is preferable to the far smaller confidence interval provided by a less variable but more biased estimator such as Chao-Wang-Jost that would probably not contain the actual values of low-order diversity (Figure 14e).

## 5. Conclusion

I have tried to evaluate the performance of diversity estimation in real conditions with simulation studies covering a reasonable set of models. Unsurprisingly, estimating diversity is more difficult when the species distribution has a heavier tail and the number of species is greater. As of the state of the art, the recommendation is to apply the Chao-Wang-Jost, the iChao-unveiled or the jackknife-unveiled estimator and consider diversity of order lower than 0.5 with caution.

When the sampling effort is high enough to allow a correct estimation of the number of species with the Chao1 estimator, the estimation by Chao-Wang-Jost is quite good down to $q = 0$. If this is not the case, the jackknife-unveiled estimator provides better results but with a higher variability. A conservative compromise for a first estimation of diversity, before choosing between Chao-Wang-Jost and jackknife-unveiled, is the iChao-unveiled estimator.

The *entropart* package for R allows computing species-neutral diversity and phylodiversity with all the estimators presented here.

## Acknowledgments

This work has benefited from an "Investissement d'Avenir" grant managed by Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-25-01).

I wish to thank the participants of the Anaee France Biodiversity group meeting in Moulis (February 2015) for fruitful discussions that motivated this paper.

## References

Ashbridge J, Goudie IBJ (2000). "Coverage-adjusted estimators for mark-recapture in heterogeneous populations." *Communications in Statistics - Simulation and Computation*, **29**(4), 1215–1237.

Beck C (2009). "Generalised Information and Entropy Measures in Physics." *Contemporary Physics*, **50**(4), 495–510.

Beck J, Schwanghart W (2010). "Comparing measures of species diversity from incomplete inventories: an update." *Methods in Ecology and Evolution*, **1**(1), 38–44.

Bonachela JA, Hinrichsen H, Muñoz MA (2008). "Entropy estimates of small data sets." *Journal of Physics A: Mathematical and Theoretical*, **41**(202001), 1–9.

Brose U, Martinez ND, Williams RJ (2003). "Estimating species richness: Sensitivity to sample coverage and insensitivity to spatial patterns." *Ecology*, **84**(9), 2364–2377.

Bulmer MG (1974). "On Fitting the Poisson Lognormal Distribution to Species-Abundance Data." *Biometrics*, **30**(1), 101–110.

Burnham KP, Overton WS (1979). "Robust Estimation of Population Size When Capture Probabilities Vary Among Animals." *Ecology*, **60**(5), 927–936.

Cao L, Grabchak M (2014). "EntropyEstimation: Estimation of Entropy and Related Quantities." URL http://cran.r-project.org/package=EntropyEstimation.

Chao A (1984). "Nonparametric estimation of the number of classes in a population." *Scandinavian Journal of Statistics*, **11**, 265–270.

Chao A, Hsieh TC, Chazdon RL, Colwell RK, Gotelli NJ (2015). "Unveiling the Species-Rank Abundance Distribution by Generalizing Good-Turing Sample Coverage Theory." *Ecology*, **96**(5), 1189–1201.

Chao A, Jost L (2012). "Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size." *Ecology*, **93**(12), 2533–2547.

Chao A, Jost L (2015). "Estimating diversity and entropy profiles via discovery rates of new species." *Methods in Ecology and Evolution*, **6**(8), 873–882.

Chao A, Lee SM, Chen TC (1988). "A generalized Good's nonparametric coverage estimator." *Chinese Journal of Mathematics*, **16**, 189–199.

Chao A, Shen TJ (2003). "Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample." *Environmental and Ecological Statistics*, **10**(4), 429–443.

Chao A, Shen TJ (2010). "Program SPADE: Species Prediction And Diversity Estimation. Program and user's guide." URL http://chao.stat.nthu.edu.tw/softwareCE.html.

Chao A, Wang YT, Jost L (2013). "Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species." *Methods in Ecology and Evolution*, **4**(11), 1091–1100.

Chiu CH, Wang YT, Walther BA, Chao A (2014). "An Improved Nonparametric Lower Bound of Species Richness via a Modified Good-Turing Frequency Formula." *Biometrics*, **70**(3), 671–682.

Cormack RM (1989). "Log-Linear Models for Capture-Recapture." *Biometrics*, **45**(2), 395–413.

Daróczy Z (1970). "Generalized information functions." *Information and Control*, **16**(1), 36–51.

Dauby G, Hardy OJ (2012). "Sampled-based estimation of diversity sensu stricto by transforming Hurlbert diversities into effective number of species." *Ecography*, **35**(7), 661–672.

Dewar RC, Porté A (2008). "Statistical mechanics unifies different ecological patterns." *Journal of theoretical biology*, **251**(3), 389–403.

Engen S, Lande R (1996). "Population dynamic models generating the lognormal species abundance distribution." *Mathematical Biosciences*, **132**(2), 169–183.

Esty WW (1983). "A Normal Limit Law for a Nonparametric Estimator of the Coverage of a Random Sample." *The Annals of Statistics*, **11**(3), 905–912.

Good IJ (1953). "On the Population Frequency of Species and the Estimation of Population Parameters." *Biometrika*, **40**(3/4), 237–264.

Gourlet-Fleury S, Guehl JM, Laroussinie O (2004). *Ecology & Management of a Neotropical Rainforest. Lessons Drawnfrom Paracou, a Long-Term Experimental Research Site in French Guiana*. Elsevier, Paris, France.

Grassberger P (1988). "Finite sample corrections to entropy and dimension estimates." *Physics Letters A*, **128**(6-7), 369–373.

Haegeman B, Hamelin J, Moriarty J, Neal P, Dushoff J, Weitz JS (2013). "Robust estimation of microbial diversity in theory and in practice." *The ISME journal*, **7**(6), 1092–101.

Havrda J, Charvát F (1967). "Quantification method of classification processes. Concept of structural a-entropy." *Kybernetika*, **3**(1), 30–35.

Hill MO (1973). "Diversity and Evenness: A Unifying Notation and Its Consequences." *Ecology*, **54**(2), 427–432.

Horvitz DG, Thompson DJ (1952). "A generalization of sampling without replacement from a finite universe." *Journal of the American Statistical Association*, **47**(260), 663–685.

Hubbell SP, Condit R, Foster RB (2005). "Barro Colorado Forest Census Plot Data." URL https://ctfs.arnarb.harvard.edu/webatlas/datasets/bci.

Jost L (2006). "Entropy and diversity." *Oikos*, **113**(2), 363–375.

Lande R, DeVries PJ, Walla TR (2000). "When species accumulation curves intersect: implications for ranking diversity using small samples." *Oikos*, **89**(3), 601–605.

Legendre P (2014). *lmodel2: Model II Regression*. R package version 1.7-2, URL http://CRAN.R-project.org/package=lmodel2.

Marcon E, Hérault B (2015a). "Decomposing Phylodiversity." *Methods in Ecology and Evolution*, **6**(3), 333–339.

Marcon E, Hérault B (2015b). "entropart, an R Package to Partition Diversity." *Journal of Statistical Software*, **67**(8), 1–26.

Marcon E, Hérault B, Baraloto C, Lang G (2012). "The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity." *Oikos*, **121**(4), 516–522.

Marcon E, Scotti I, Hérault B, Rossi V, Lang G (2014a). "Generalization of the partitioning of Shannon diversity." *Plos One*, **9**(3), e90289.

Marcon E, Zhang Z, Hérault B (2014b). "The Decomposition of Similarity-Based Diversity and its Bias Correction." *HAL*, **00989454**(version 3), 1–10.

Motomura I (1932). "On the statistical treatment of communities." *Zoological Magazine*, **44**, 379–383.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H (2012). "vegan: Community Ecology Package." URL http://cran.r-project.org/package=vegan.

Patil GP, Taillie C (1982). "Diversity as a concept and its measurement." *Journal of the American Statistical Association*, **77**(379), 548–561.

Preston FW (1948). "The Commonness, And Rarity, of Species." *Ecology*, **29**(3), 254–283.

Pueyo S, He F, Zillio T (2007). "The maximum entropy formalism and the idiosyncratic theory of biodiversity." *Ecology letters*, **10**(11), 1017–28.

R Development Core Team (2015). "R: A Language and Environment for Statistical Computing." URL http://www.r-project.org.

Tokeshi M (1990). "Niche Apportionment or Random Assortment: Species Abundance Patterns Revisited." *Journal of Animal Ecology*, **59**(3), 1129–1146.

Tokeshi M (1993). "Species Abundance Patterns and Community Structure." *Advances in Ecological Research*, **24**, 111–186.

Tothmeresz B (1995). "Comparison of different methods for diversity ordering." *Journal of Vegetation Science*, **6**(2), 283–290.

Tsallis C (1988). "Possible generalization of Boltzmann-Gibbs statistics." *Journal of Statistical Physics*, **52**(1), 479–487.

Tsallis C (1994). "What are the numbers that experiments provide?" *Química Nova*, **17**(6), 468–471.

Volkov I, Banavar JR, Hubbell SP, Maritan A (2003). "Neutral theory and relative species abundance in ecology." *Nature*, **424**(6952), 1035–1037.

Wang JP (2011). "SPECIES: An R Package for Species Richness Estimation." *Journal of Statistical Software*, **40**(9), 1–15.

Whittaker RH (1972). "Evolution and Measurement of Species Diversity." *Taxon*, **21**(2/3), 213–251.

Williamson M, Gaston KJ (2005). "The lognormal distribution is not an appropriate null hypothesis for the species-abundance distribution." *Journal of Animal Ecology*, **74**(2001), 409–422.

Zhang Z (2013). "Asymptotic normality of an entropy estimator with exponentially decaying bias." *IEEE Transactions on Information Theory*, **59**(1), 504–508.

Zhang Z, Grabchak M (2013). "Bias adjustment for a nonparametric entropy estimator." *Entropy*, **15**(6), 1999–2011.

Zhang Z, Grabchak M (2014). "Entropic Representation and Estimation of Diversity Indices." *arXiv*, **1403.3031**(v. 2), 1–12.

Zhang Z, Huang H (2007). "Turing's formula revisited." *Journal of Quantitative Linguistics*, **14**(2-3), 222–241.

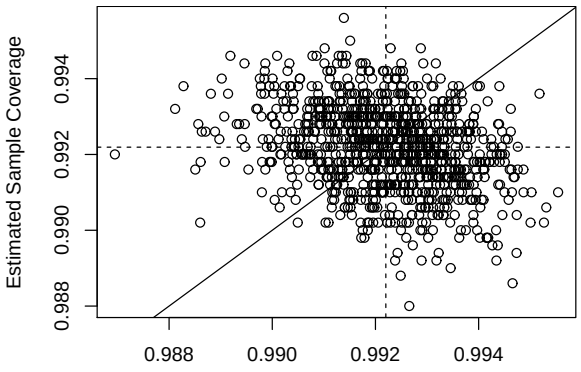
Zhang Z, Zhou J (2010). "Re-parameterization of multinomial distributions and diversity indices." *Journal of Statistical Planning and Inference*, **140**(7), 1731–1738.

## Appendix 1: Sample coverage estimation

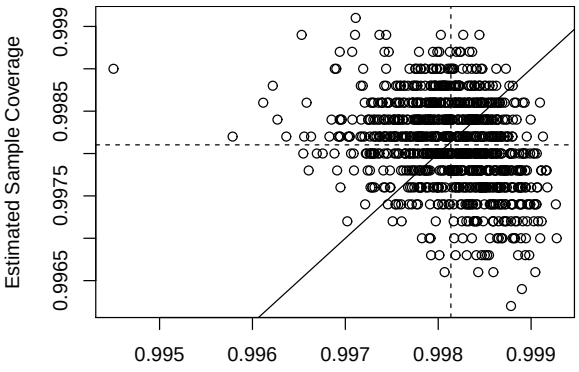
Figures 7 and 8 compares the estimated and the real sample coverages of 1000 samples of sizes between 200 and 5000 individuals from a lognormal and a geometric distribution of 300 or 600 species The average estimated sample coverage virtually equals the average real coverage even when the sampling size is small.

## Appendix 2: Estimated diversity profiles

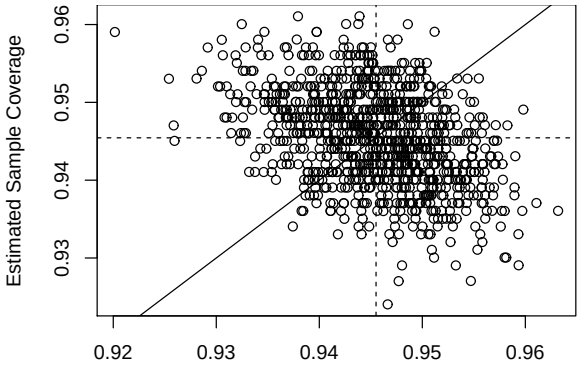
Figures 9 to 17 show the estimation of diversity profiles of communities of 300 species. Each figure presents an estimator. The diversity of the lognormal and of the geometric community is estimated, for sample sizes from 200 to 5000 individuals. 1000 simulations were performed to produce a 95% confidence envelope of the profiles. Figures 18 and **??** present the same profiles for 600-species communities, limited to the best two estimators.



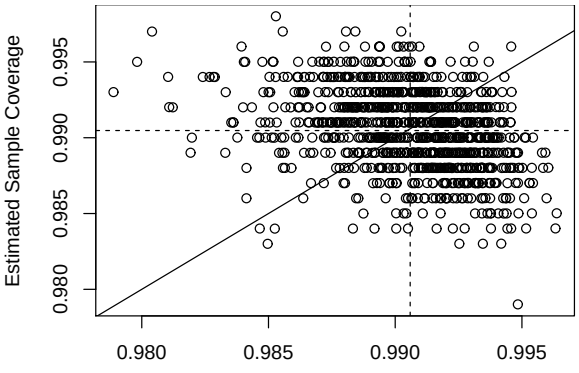

**(a)** Lognormal, 5000 individuals



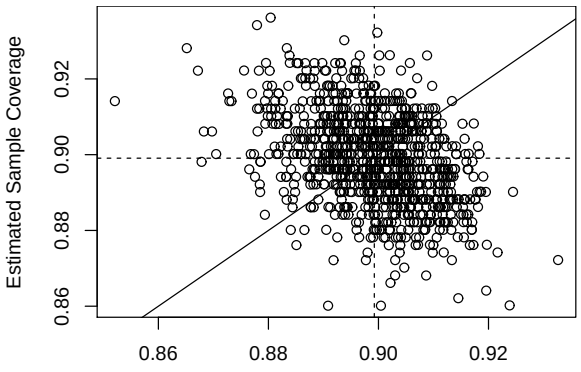

**(b)** Geometric, 5000 individuals



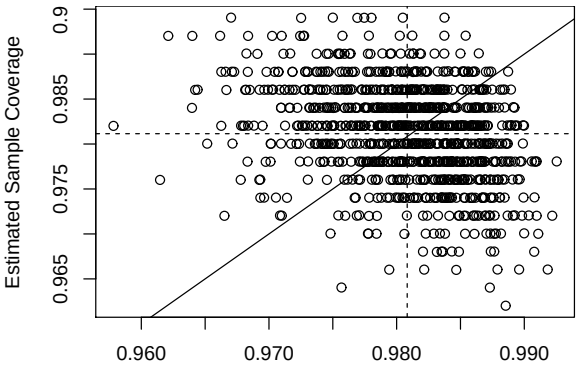

**(c)** Lognormal, 1000 individuals



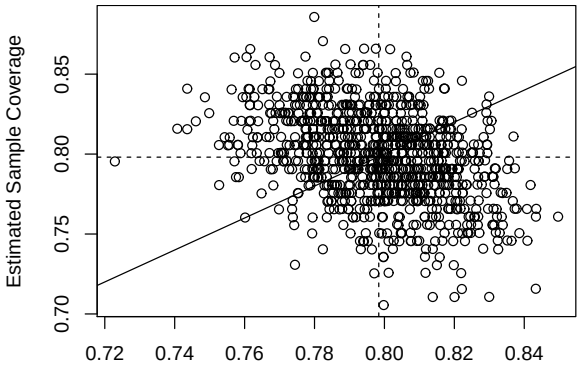

**(d)** Geometric, 1000 individuals



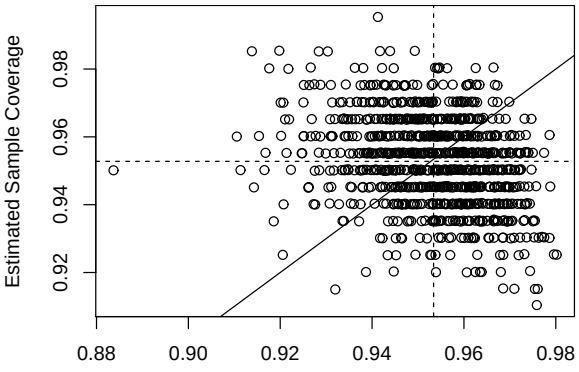

**(e)** Lognormal, 500 individuals




**(f)** Geometric, 500 individuals

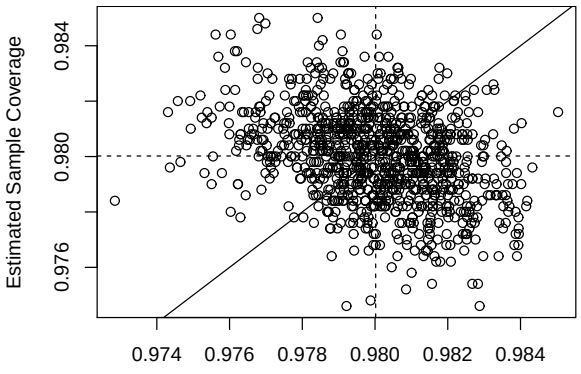



**(g)** Lognormal, 200 individuals

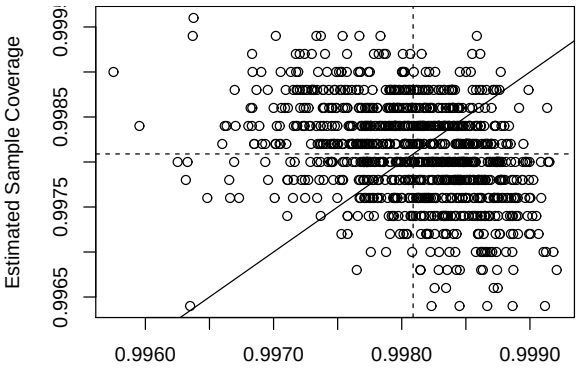



**(h)** Geometric, 200 individuals

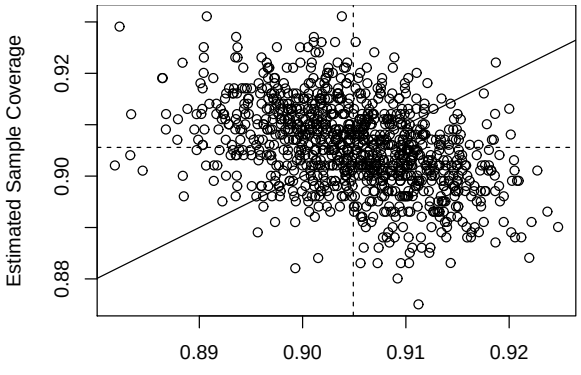
**Figure 7.** Estimated *vs* real sample coverage of simulated samples. The dotted lines are the average values. The distributions contain 300 species. The plain line represents equality.



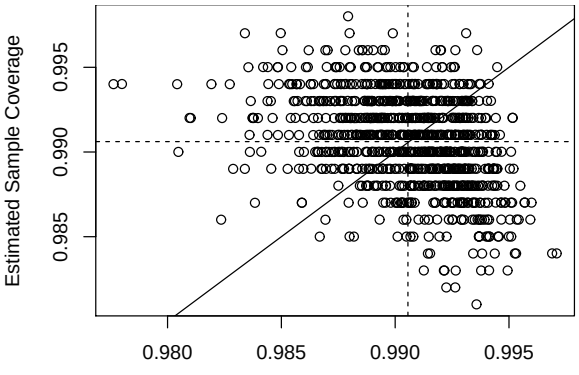

**(a)** Lognormal, 5000 individuals



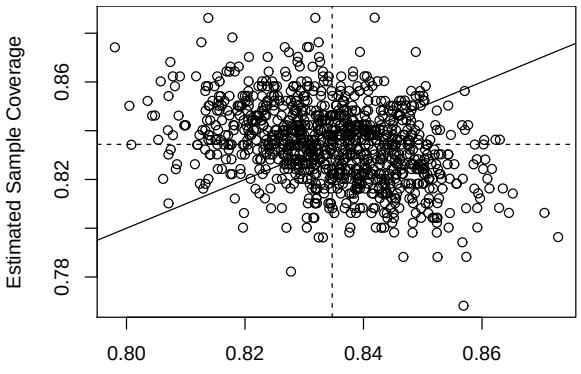

**(b)** Geometric, 5000 individuals



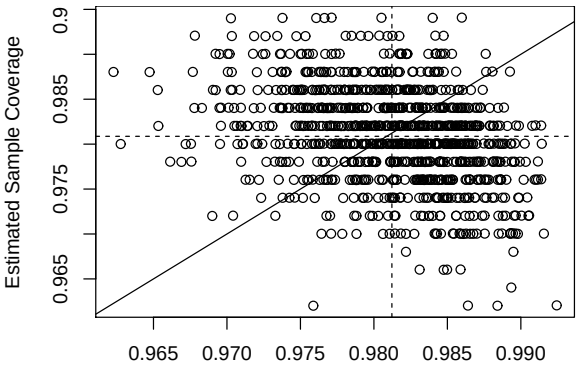

**(c)** Lognormal, 1000 individuals



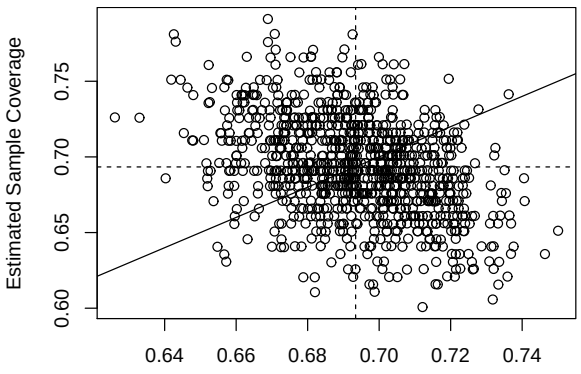

**(d)** Geometric, 1000 individuals



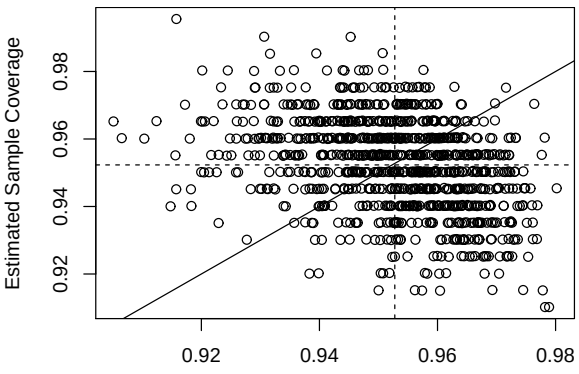

**(e)** Lognormal, 500 individuals




**(f)** Geometric, 500 individuals

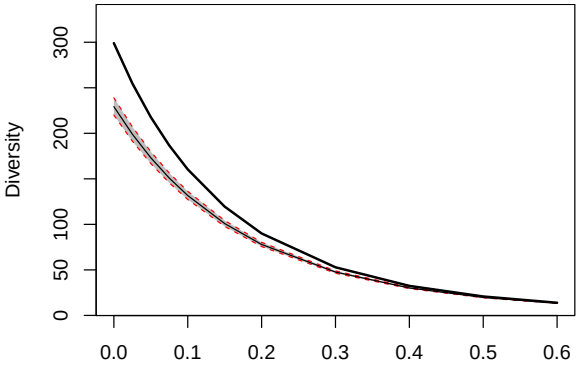



**(g)** Lognormal, 200 individuals

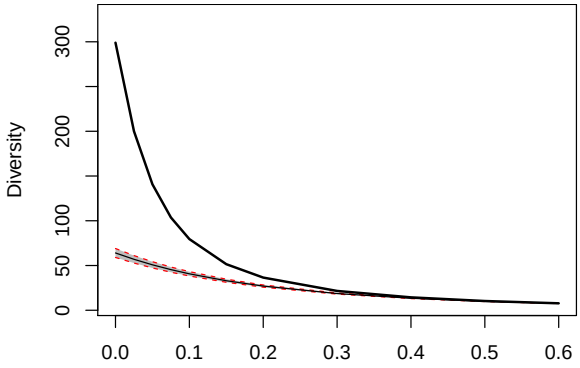



**(h)** Geometric, 200 individuals

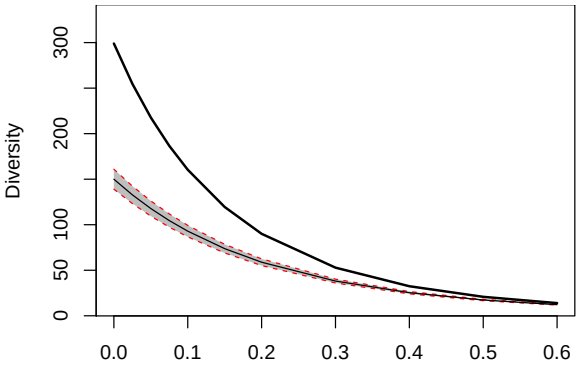
**Figure 8.** Estimated *vs* real sample coverage of simulated samples. The dotted lines are the average values. The distributions contain 600 species. The plain line represents equality.



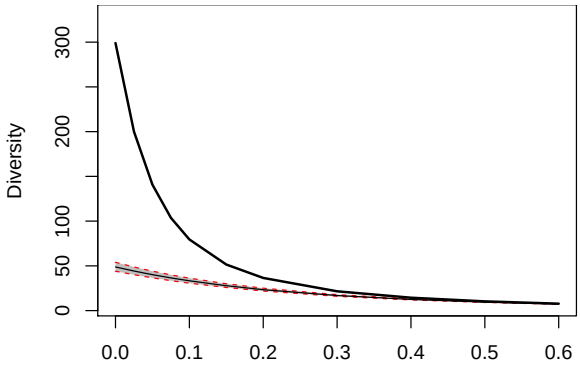

**(a)** Lognormal, 300 species, 5000 individuals



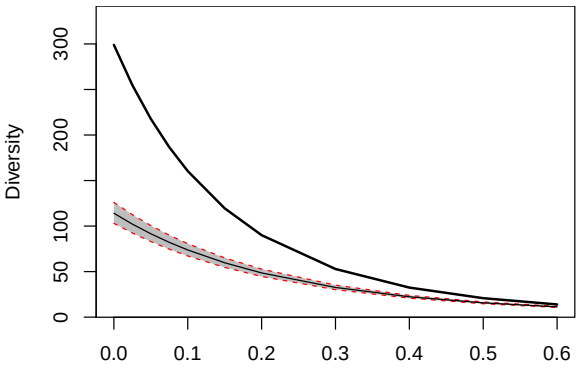

**(b)** Geometric, 300 species, 5000 individuals



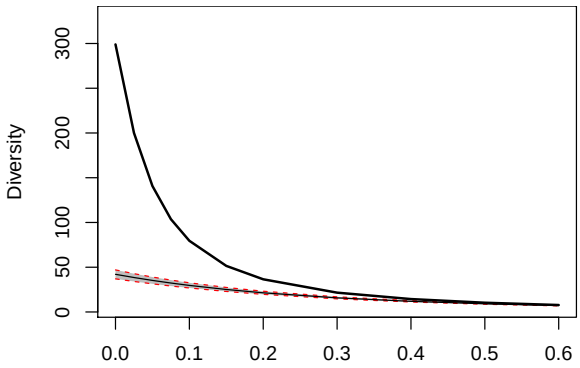

**(c)** Lognormal, 300 species, 1000 individuals



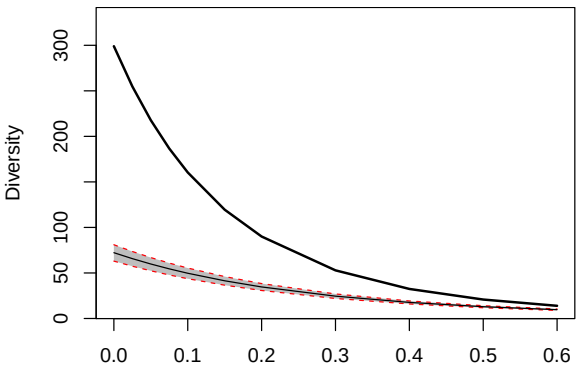

**(d)** Geometric, 300 species, 1000 individuals



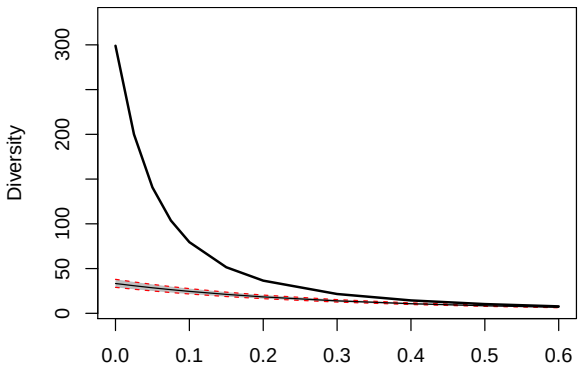

**(e)** Lognormal, 300 species, 500 individuals




**(f)** Geometric, 300 species, 500 individuals

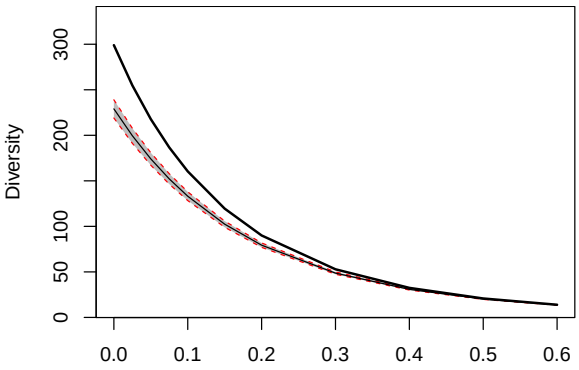



**(g)** Lognormal, 300 species, 200 individuals

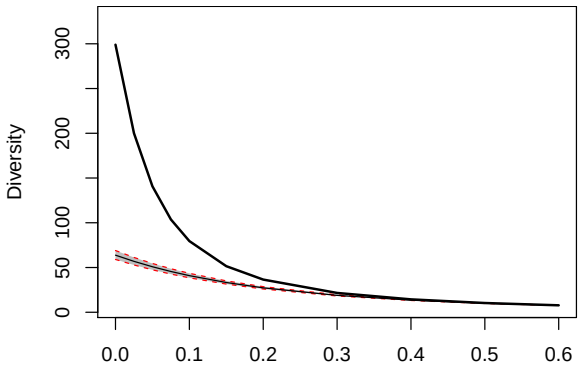



**(h)** Geometric, 300 species, 200 individuals

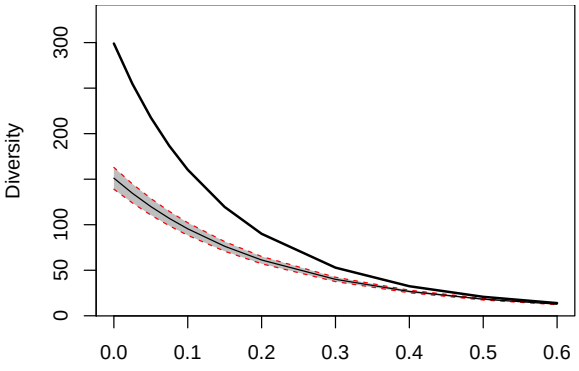
**Figure 9.** Estimation by the plug-in estimator of the diversity profiles of simulated lognormal (left) and geometric (right) communities. The sample size decreases from 5000 (top) to 200 (bottom) individuals. The 95% confidence envelope of the estimation is shaded. The real diversity is plotted by the bold line.



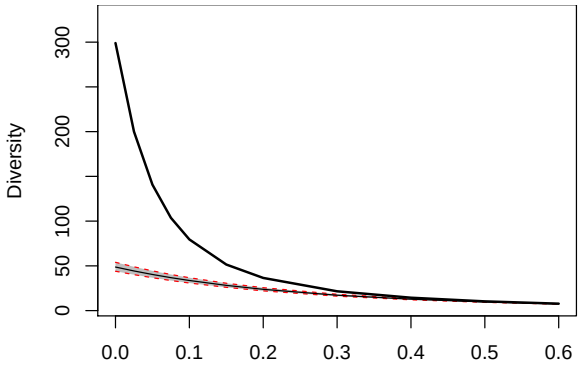

**(a)** Lognormal, 300 species, 5000 individuals



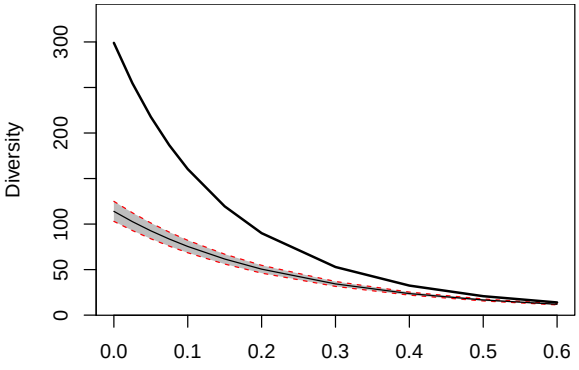

**(b)** Geometric, 300 species, 5000 individuals



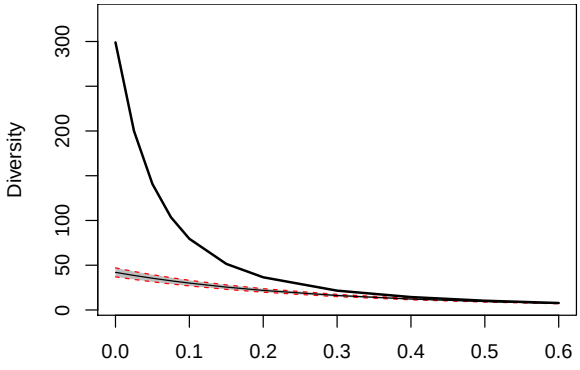

**(c)** Lognormal, 300 species, 1000 individuals



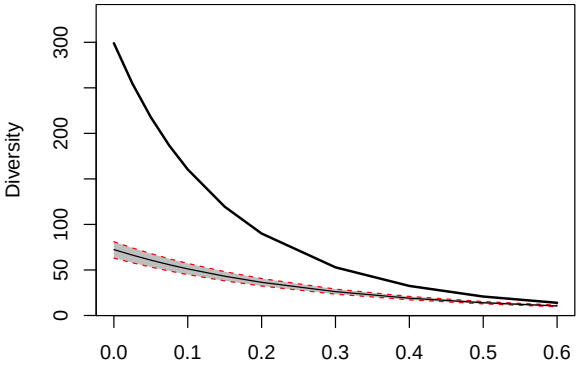

**(d)** Geometric, 300 species, 1000 individuals



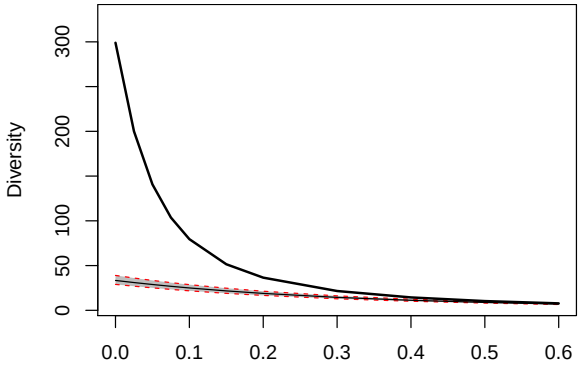

**(e)** Lognormal, 300 species, 500 individuals




**(f)** Geometric, 300 species, 500 individuals

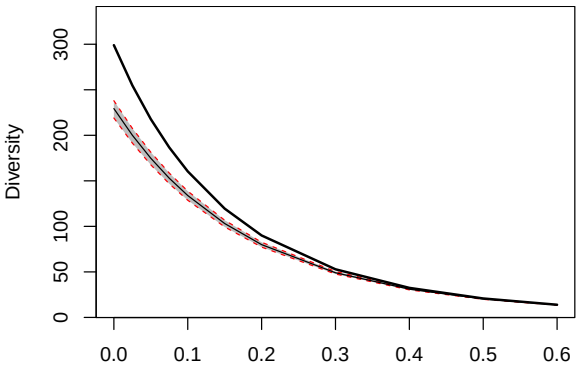



**(g)** Lognormal, 300 species, 200 individuals

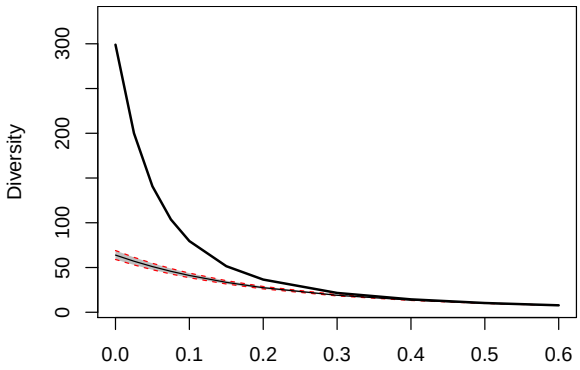



**(h)** Geometric, 300 species, 200 individuals

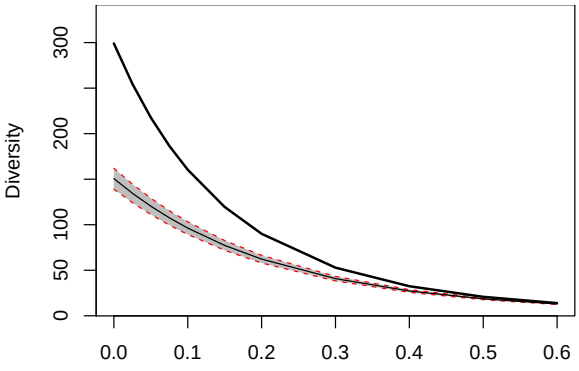
**Figure 10.** Estimation by the Zhang-Grabchak estimator of the diversity profiles of simulated lognormal (left) and geometric (right) communities. The sample size decreases from 5000 (top) to 200 (bottom) individuals. The 95% confidence envelope of the estimation is shaded. The real diversity is plotted by the bold line.



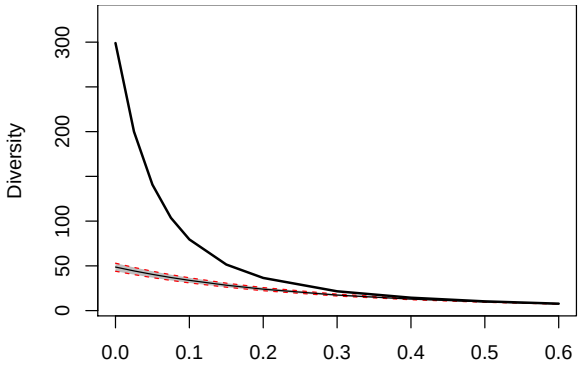

**(a)** Lognormal, 300 species, 5000 individuals



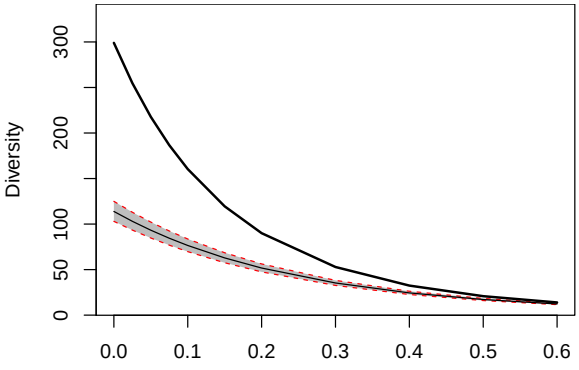

**(b)** Geometric, 300 species, 5000 individuals



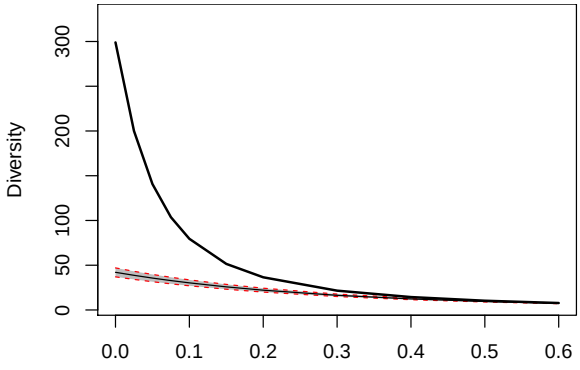

**(c)** Lognormal, 300 species, 1000 individuals



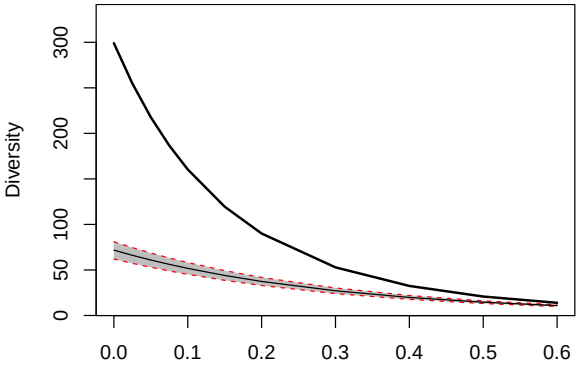

**(d)** Geometric, 300 species, 1000 individuals



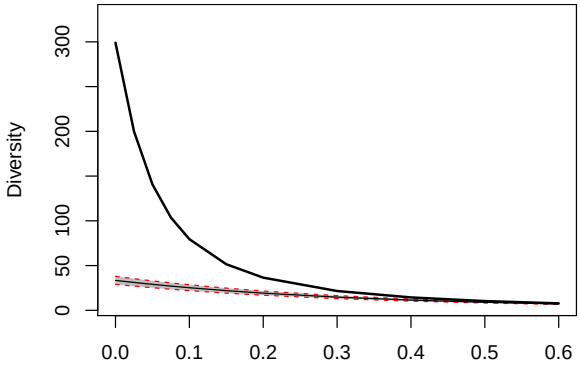

**(e)** Lognormal, 300 species, 500 individuals




**(f)** Geometric, 300 species, 500 individuals

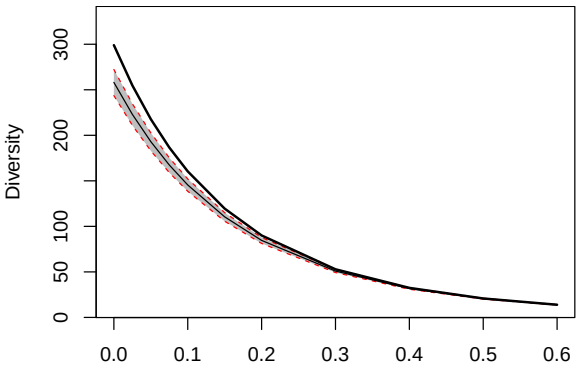



**(g)** Lognormal, 300 species, 200 individuals

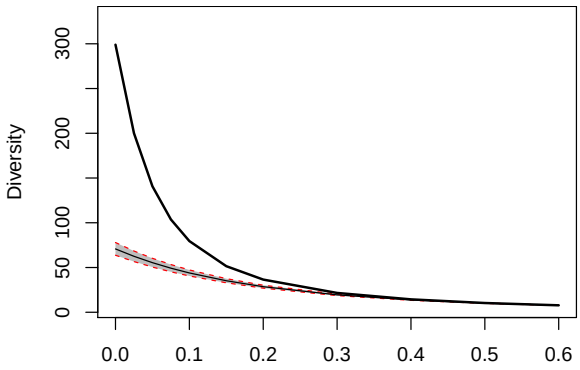



**(h)** Geometric, 300 species, 200 individuals

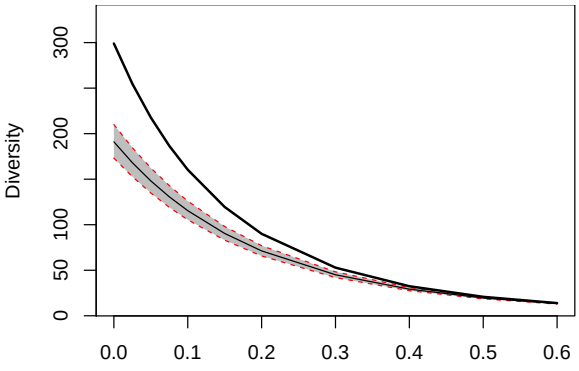
**Figure 11.** Estimation by the Grassberger estimator of the diversity profiles of simulated lognormal (left) and geometric (right) communities. The sample size decreases from 5000 (top) to 200 (bottom) individuals. The 95% confidence envelope of the estimation is shaded. The real diversity is plotted by the bold line.



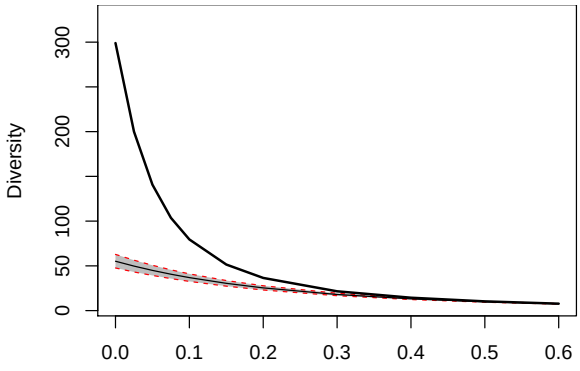

**(a)** Lognormal, 300 species, 5000 individuals



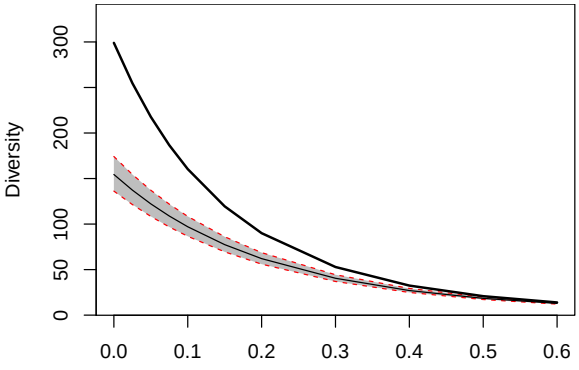

**(b)** Geometric, 300 species, 5000 individuals



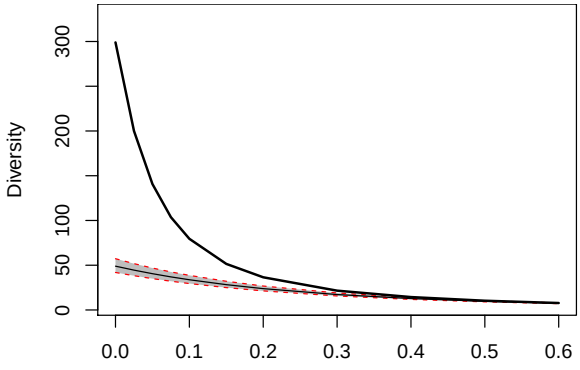

**(c)** Lognormal, 300 species, 1000 individuals



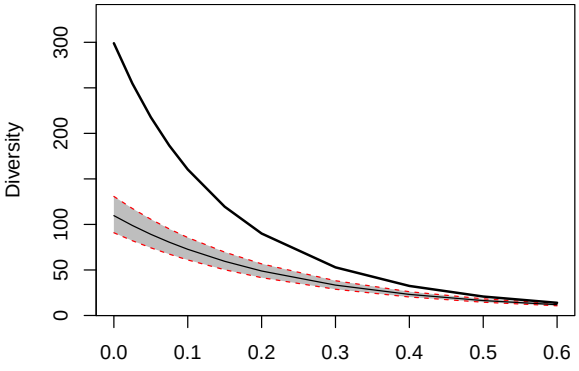

**(d)** Geometric, 300 species, 1000 individuals



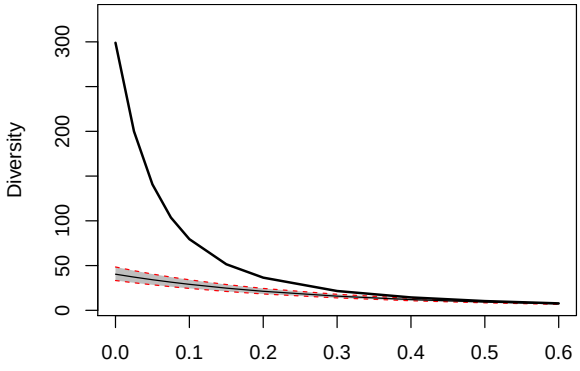

**(e)** Lognormal, 300 species, 500 individuals




**(f)** Geometric, 300 species, 500 individuals

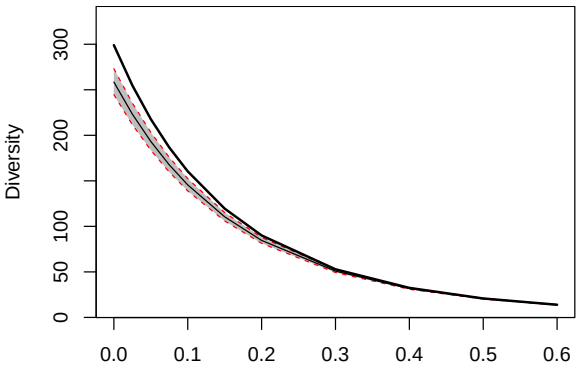



**(g)** Lognormal, 300 species, 200 individuals

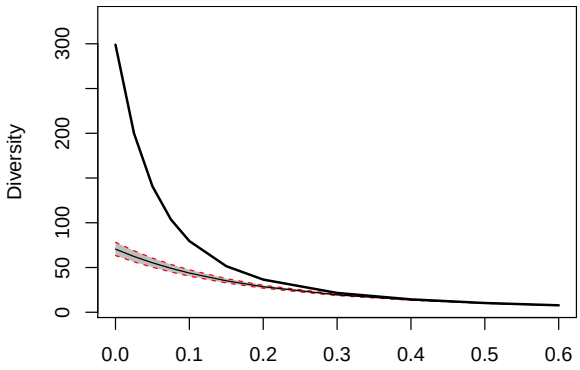



**(h)** Geometric, 300 species, 200 individuals

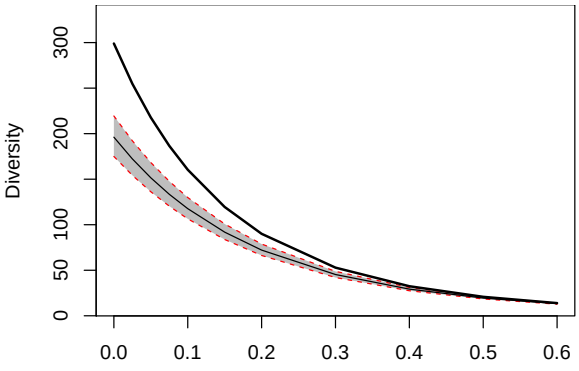
**Figure 12.** Estimation by the Chao-Shen estimator of the diversity profiles of simulated lognormal (left) and geometric (right) communities. The sample size decreases from 5000 (top) to 200 (bottom) individuals. The 95% confidence envelope of the estimation is shaded. The real diversity is plotted by the bold line.



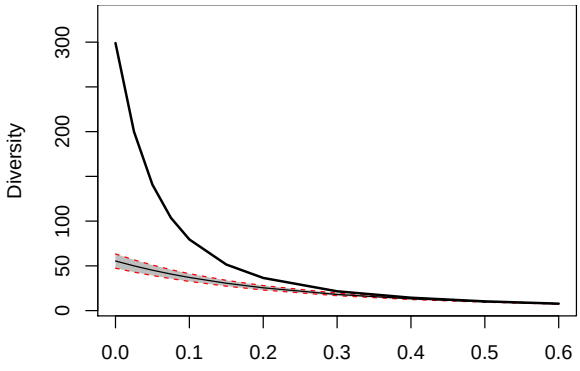

**(a)** Lognormal, 300 species, 5000 individuals



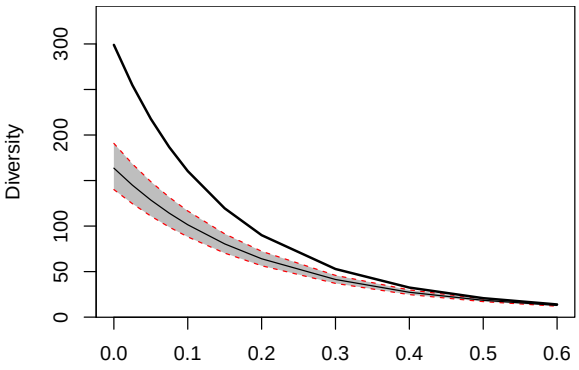

**(b)** Geometric, 300 species, 5000 individuals



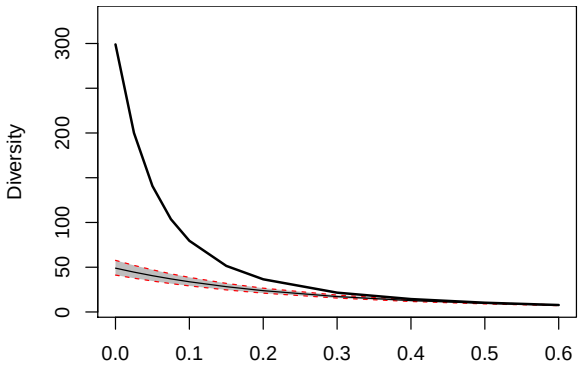

**(c)** Lognormal, 300 species, 1000 individuals



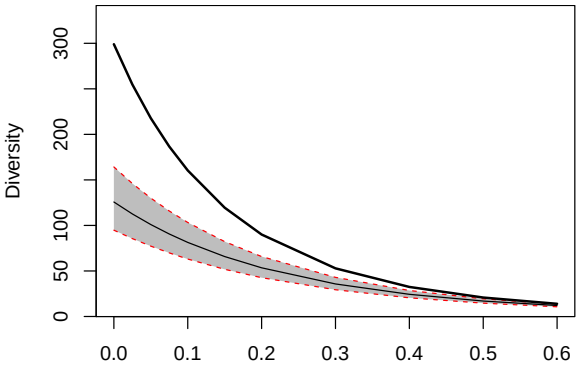

**(d)** Geometric, 300 species, 1000 individuals



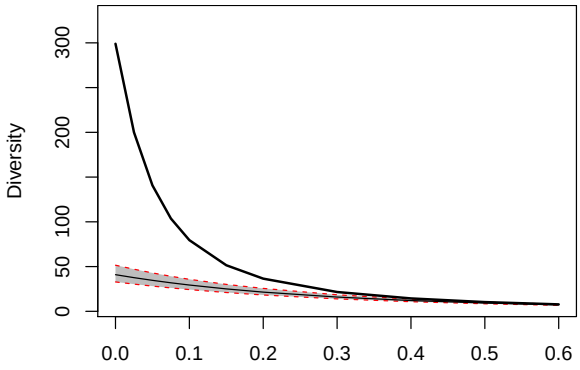

**(e)** Lognormal, 300 species, 500 individuals




**(f)** Geometric, 300 species, 500 individuals

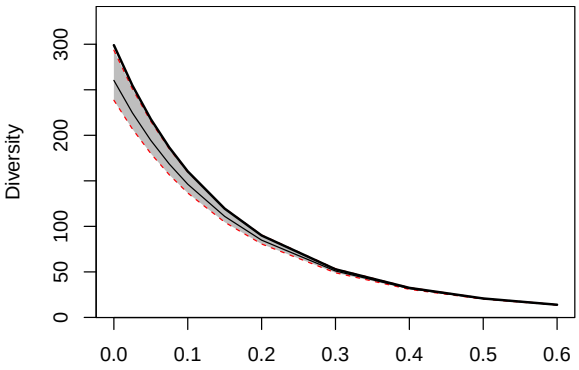



**(g)** Lognormal, 300 species, 200 individuals

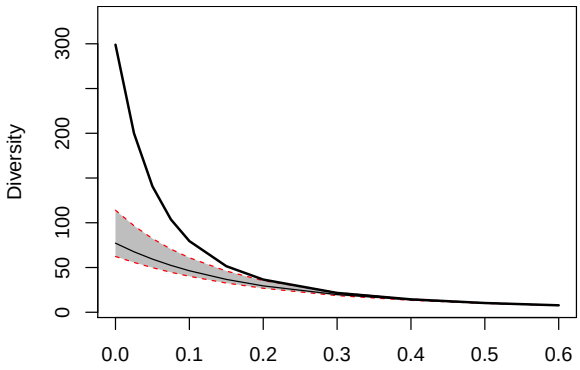



**(h)** Geometric, 300 species, 200 individuals

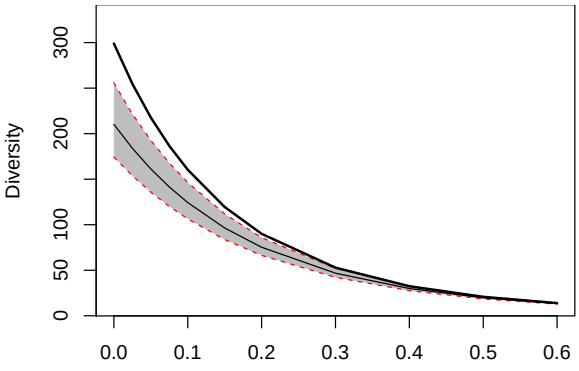
**Figure 13.** Estimation by the Generalized-Coverage estimator of the diversity profiles of simulated lognormal (left) and geometric (right) communities. The sample size decreases from 5000 (top) to 200 (bottom) individuals. The 95% confidence envelope of the estimation is shaded. The real diversity is plotted by the bold line.



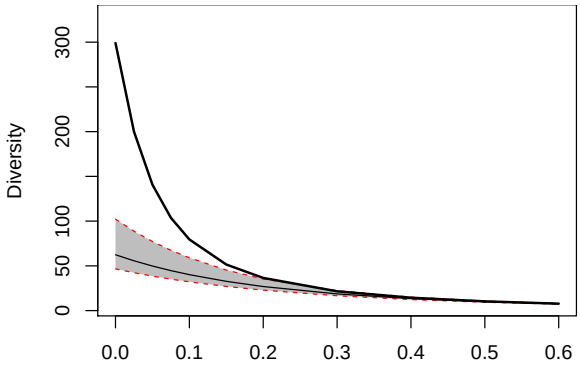

**(a)** Lognormal, 300 species, 5000 individuals



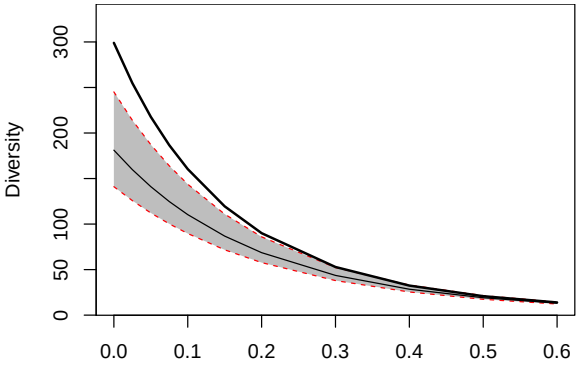

**(b)** Geometric, 300 species, 5000 individuals



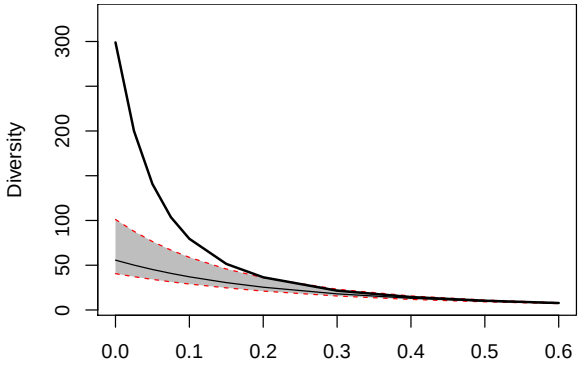

**(c)** Lognormal, 300 species, 1000 individuals



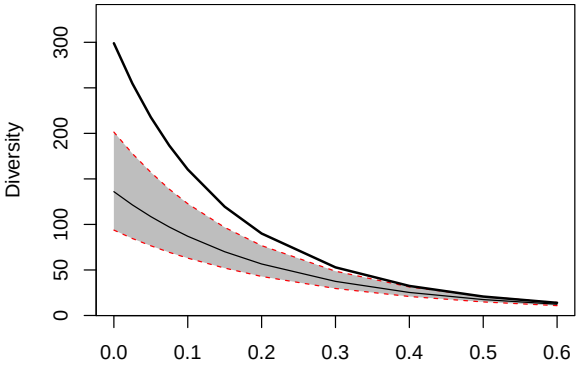

**(d)** Geometric, 300 species, 1000 individuals



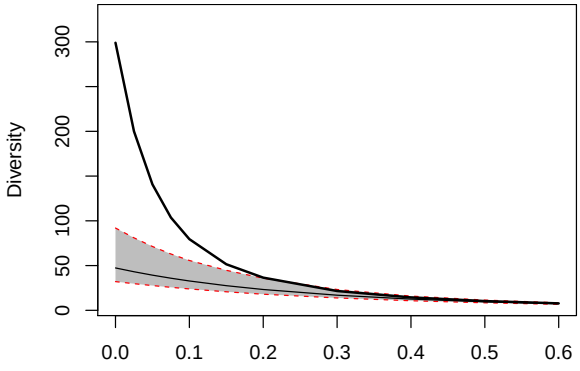

**(e)** Lognormal, 300 species, 500 individuals




**(f)** Geometric, 300 species, 500 individuals

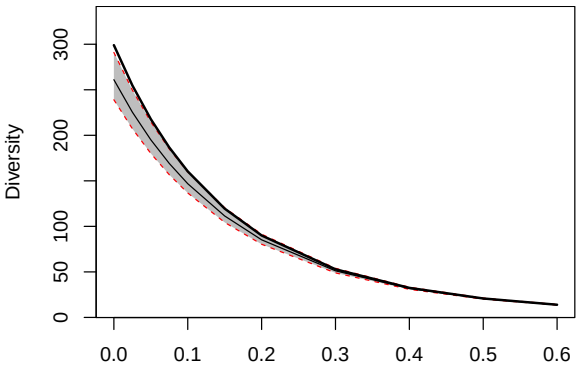



**(g)** Lognormal, 300 species, 200 individuals

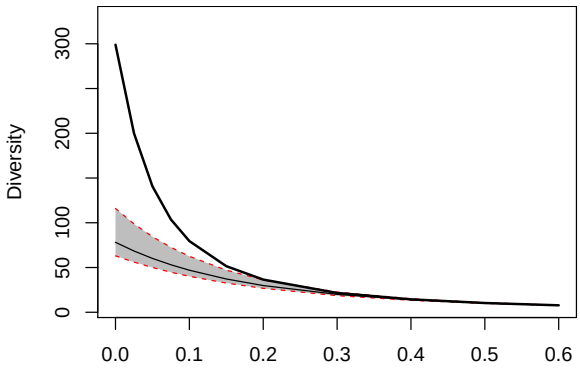



**(h)** Geometric, 300 species, 200 individuals

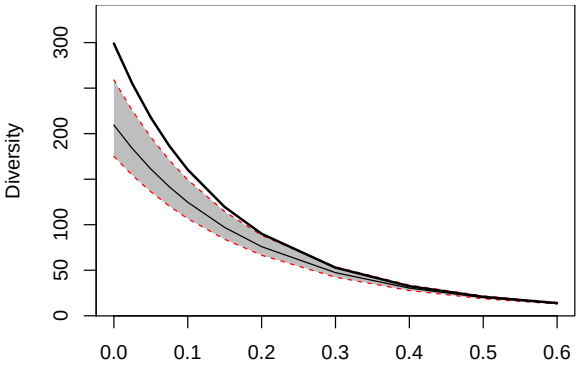
**Figure 14.** Estimation by the Chao-Wang-Jost estimator of the diversity profiles of simulated lognormal (left) and geometric (right) communities. The sample size decreases from 5000 (top) to 200 (bottom) individuals. The 95% confidence envelope of the estimation is shaded. The real diversity is plotted by the bold line.



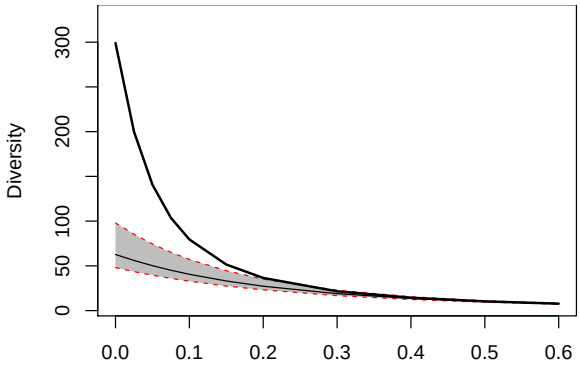

**(a)** Lognormal, 300 species, 5000 individuals



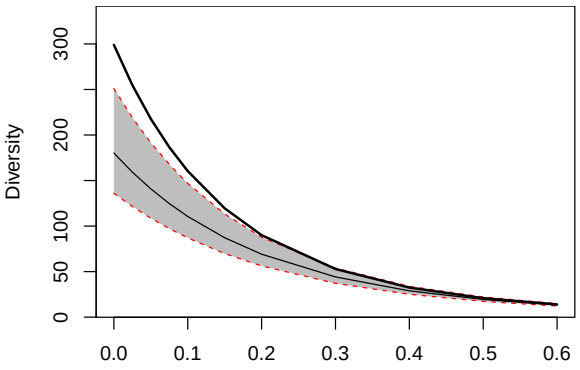

**(b)** Geometric, 300 species, 5000 individuals



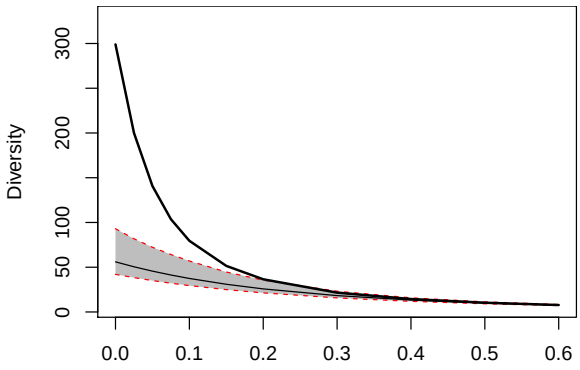

**(c)** Lognormal, 300 species, 1000 individuals



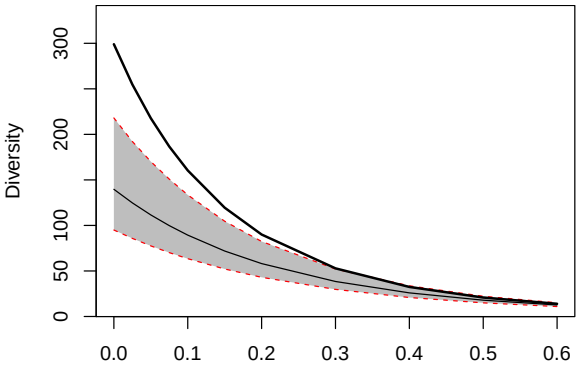

**(d)** Geometric, 300 species, 1000 individuals



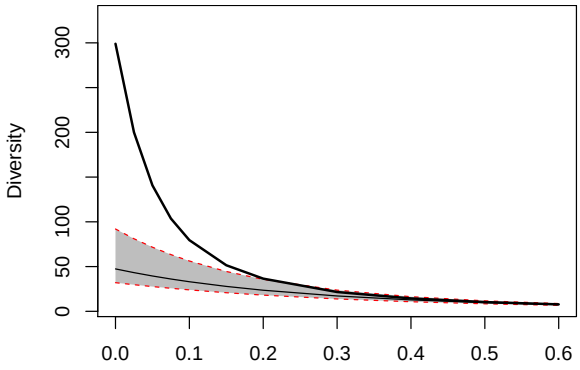

**(e)** Lognormal, 300 species, 500 individuals




**(f)** Geometric, 300 species, 500 individuals

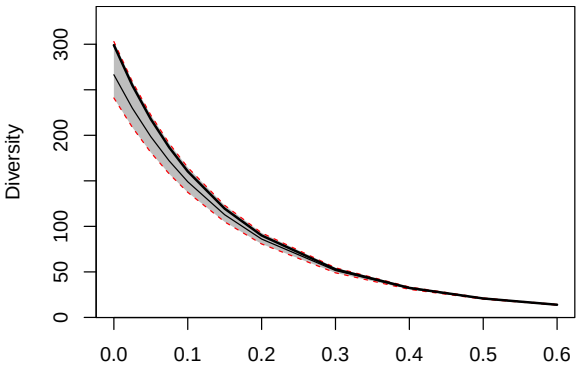



**(g)** Lognormal, 300 species, 200 individuals

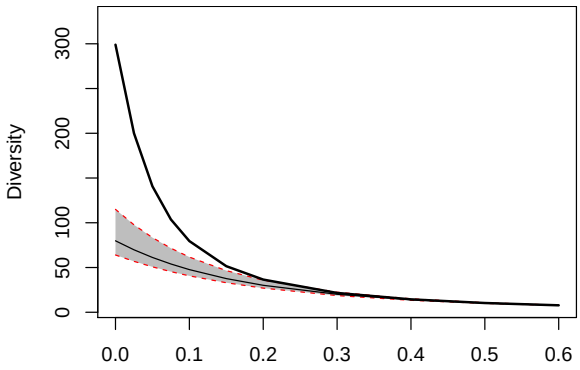



**(h)** Geometric, 300 species, 200 individuals

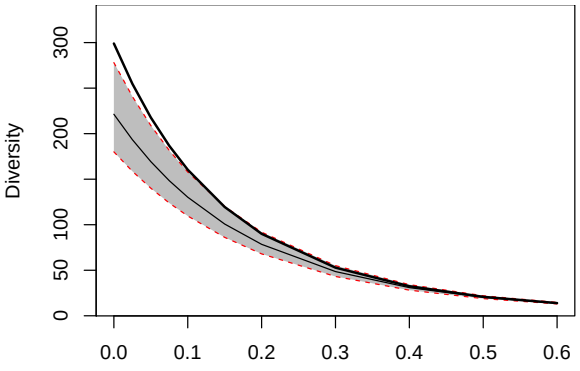
**Figure 15.** Estimation by the Chao-unveiled estimator of the diversity profiles of simulated lognormal (left) and geometric (right) communities. The sample size decreases from 5000 (top) to 200 (bottom) individuals. The 95% confidence envelope of the estimation is shaded. The real diversity is plotted by the bold line.



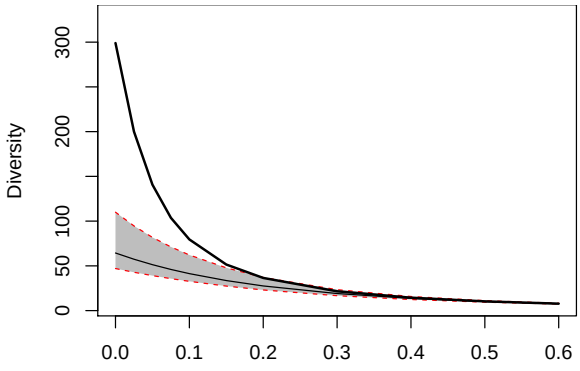

**(a)** Lognormal, 300 species, 5000 individuals



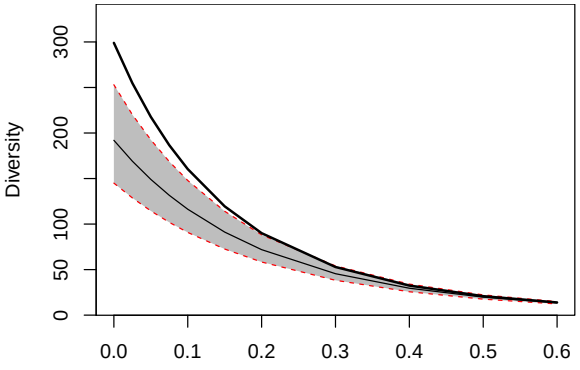

**(b)** Geometric, 300 species, 5000 individuals



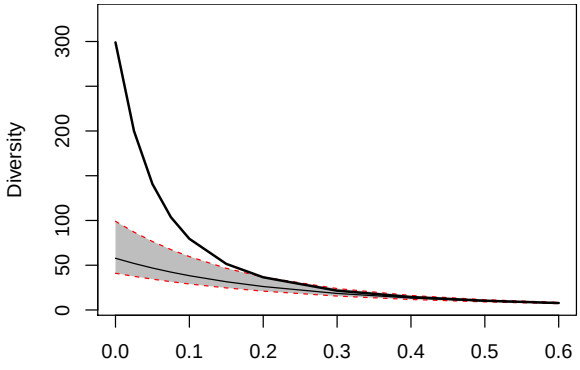

**(c)** Lognormal, 300 species, 1000 individuals



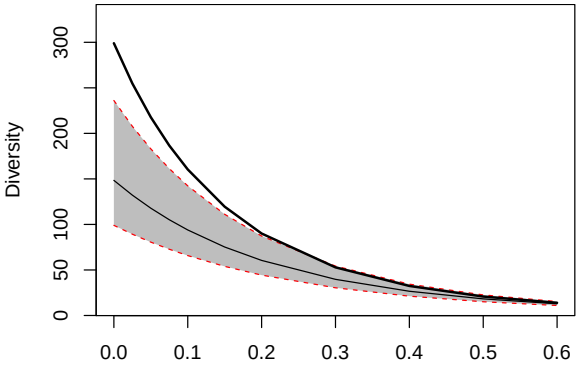

**(d)** Geometric, 300 species, 1000 individuals



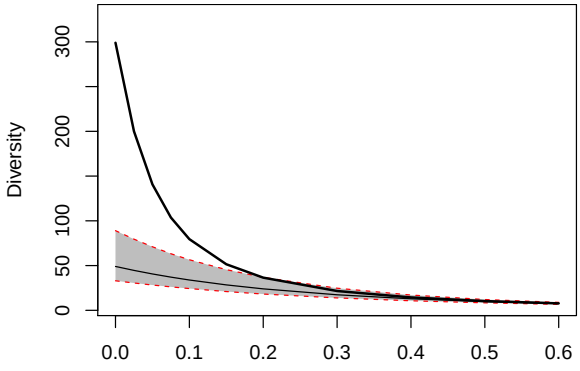

**(e)** Lognormal, 300 species, 500 individuals




**(f)** Geometric, 300 species, 500 individuals

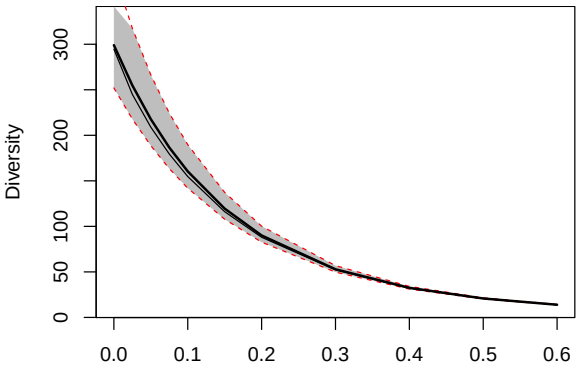



**(g)** Lognormal, 300 species, 200 individuals

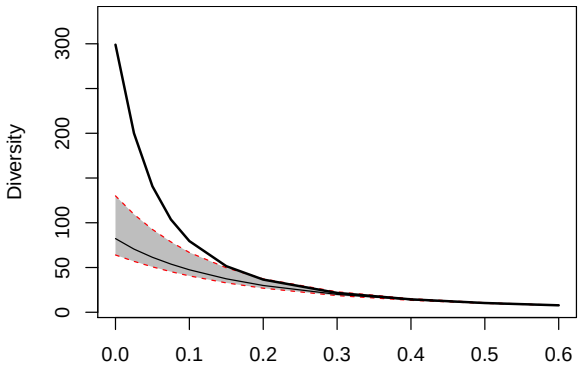



**(h)** Geometric, 300 species, 200 individuals

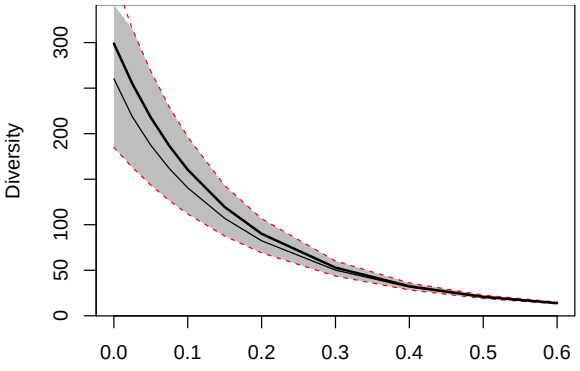
**Figure 16.** Estimation by the iChao-unveiled estimator of the diversity profiles of simulated lognormal (left) and geometric (right) communities. The sample size decreases from 5000 (top) to 200 (bottom) individuals. The 95% confidence envelope of the estimation is shaded. The real diversity is plotted by the bold line.



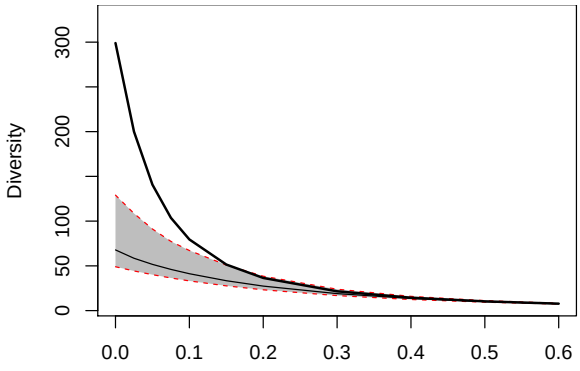

**(a)** Lognormal, 300 species, 5000 individuals



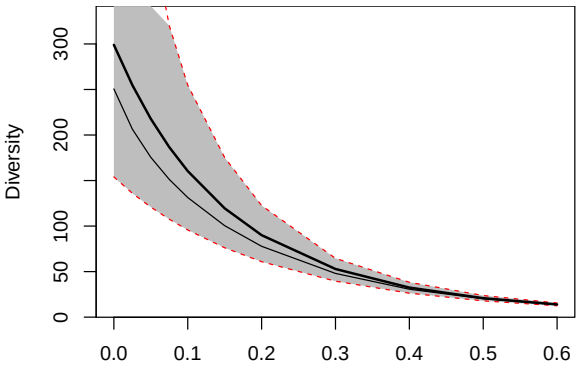

**(b)** Geometric, 300 species, 5000 individuals



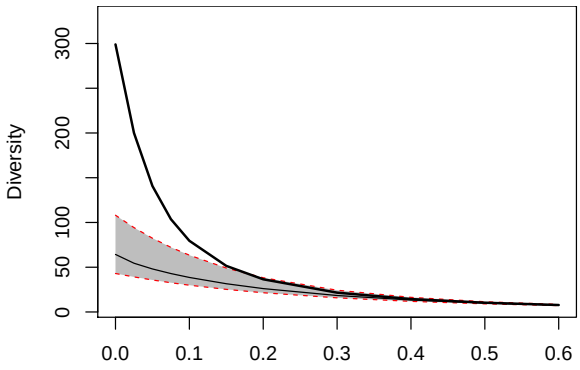

**(c)** Lognormal, 300 species, 1000 individuals



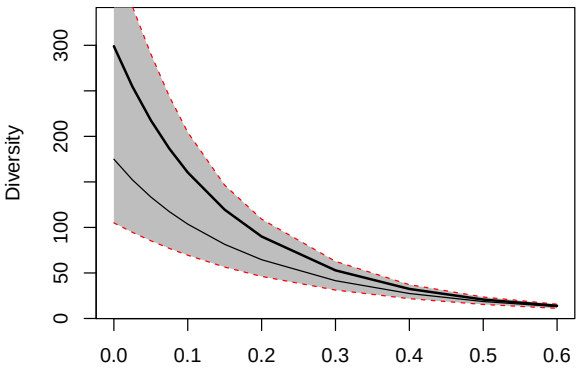

**(d)** Geometric, 300 species, 1000 individuals



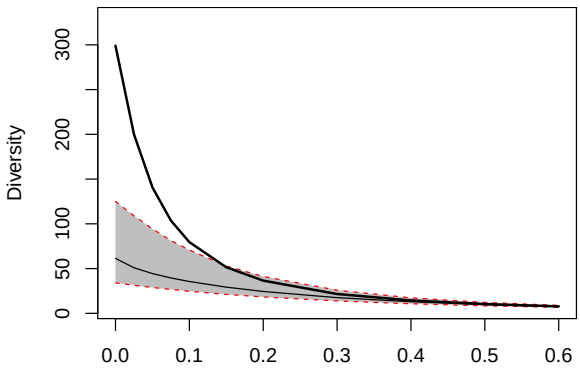

**(e)** Lognormal, 300 species, 500 individuals




**(f)** Geometric, 300 species, 500 individuals

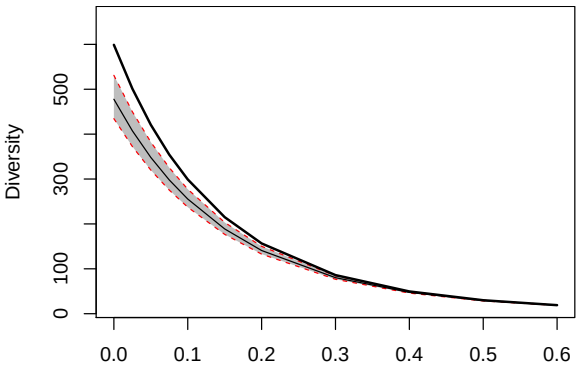



**(g)** Lognormal, 300 species, 200 individuals

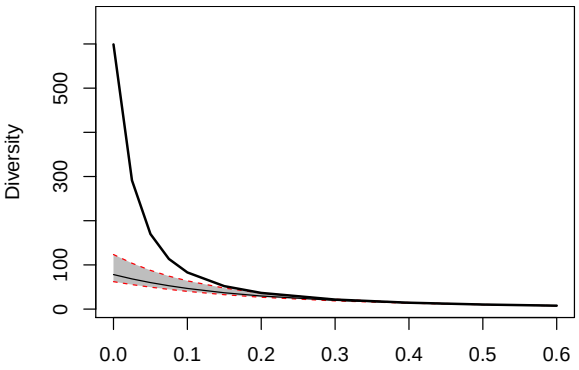



**(h)** Geometric, 300 species, 200 individuals

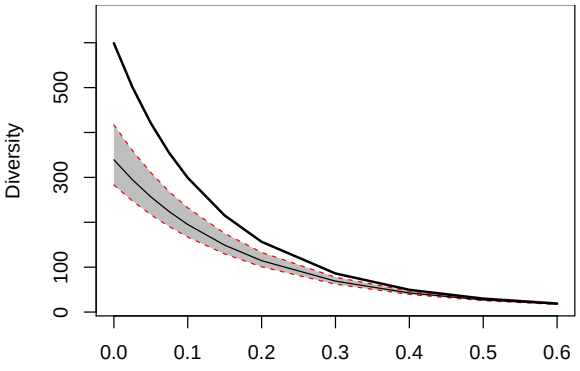
**Figure 17.** Estimation by the jackknife-unveiled estimator of the diversity profiles of simulated lognormal (left) and geometric (right) communities. The sample size decreases from 5000 (top) to 200 (bottom) individuals. The 95% confidence envelope of the estimation is shaded. The real diversity is plotted by the bold line.



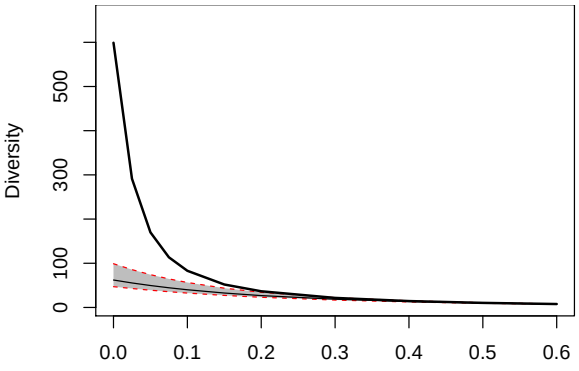

**(a)** Lognormal, 600 species, 5000 individuals



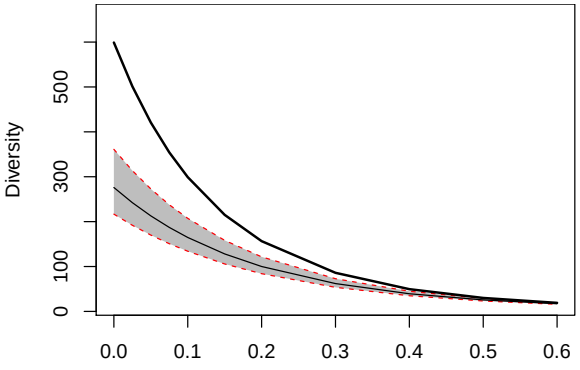

**(b)** Geometric, 600 species, 5000 individuals



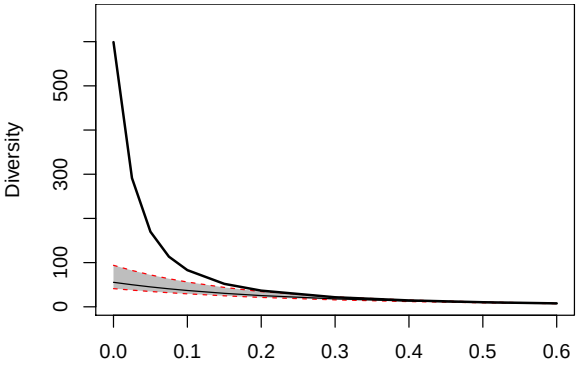

**(c)** Lognormal, 600 species, 1000 individuals



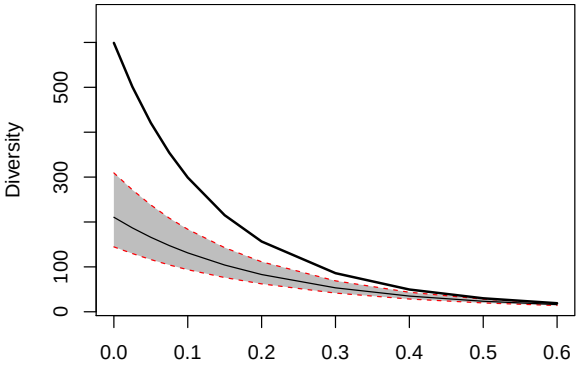

**(d)** Geometric, 600 species, 1000 individuals



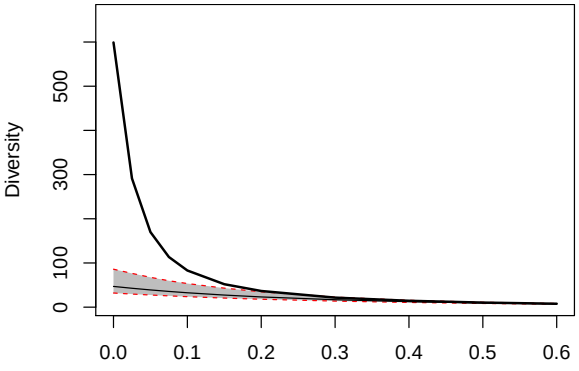

**(e)** Lognormal, 600 species, 500 individuals




**(f)** Geometric, 600 species, 500 individuals

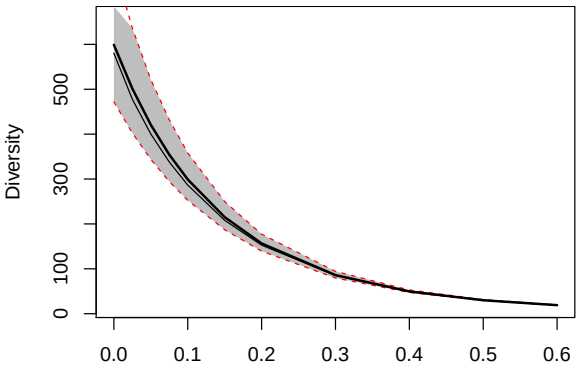



**(g)** Lognormal, 600 species, 200 individuals

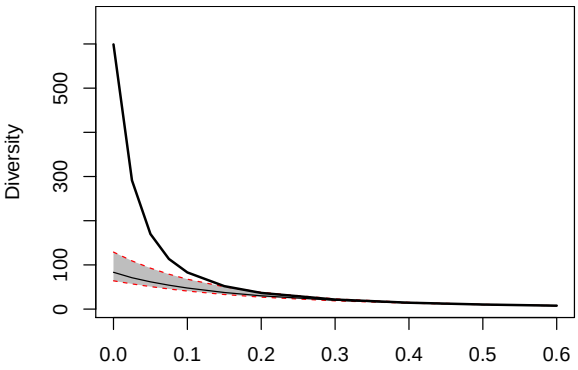



**(h)** Geometric, 600 species, 200 individuals

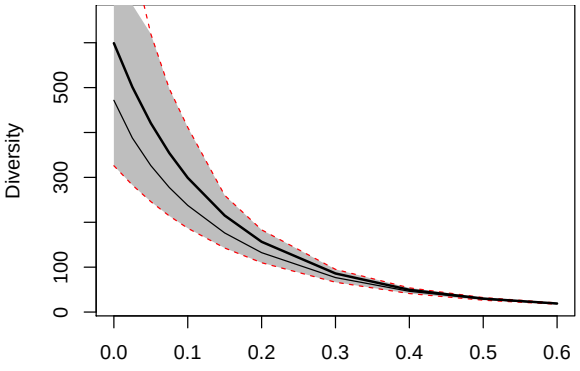
**Figure 18.** Estimation by the Chao-Wang-Jost estimator of the diversity profiles of simulated lognormal (left) and geometric (right) communities. The sample size decreases from 5000 (top) to 200 (bottom) individuals. The 95% confidence envelope of the estimation is shaded. The real diversity is plotted by the bold line.



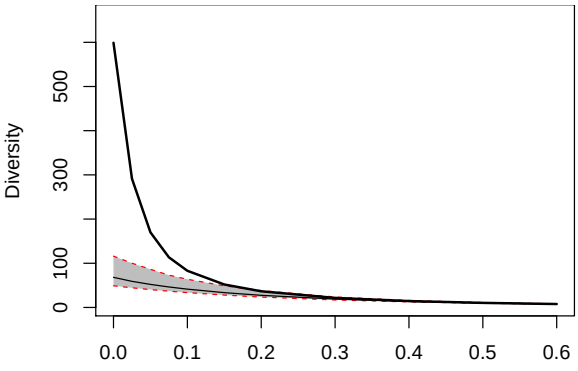

**(a)** Lognormal, 600 species, 5000 individuals



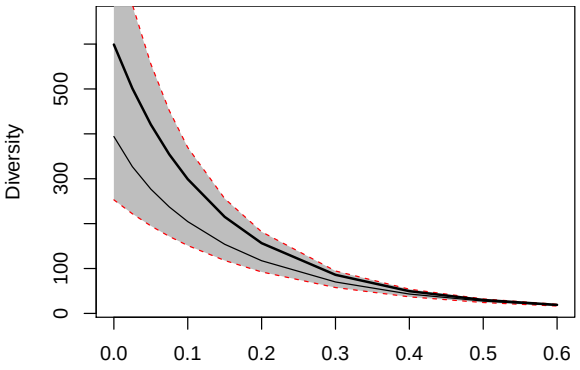

**(b)** Geometric, 600 species, 5000 individuals



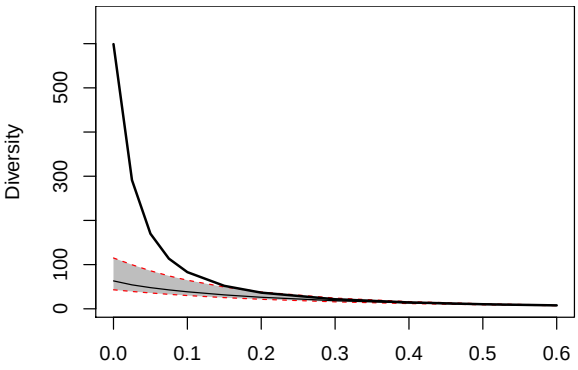

**(c)** Lognormal, 600 species, 1000 individuals



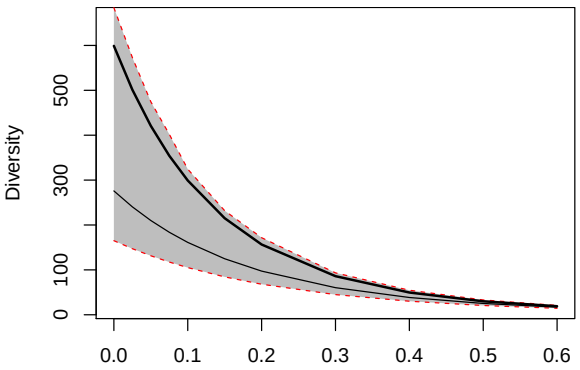

**(d)** Geometric, 600 species, 1000 individuals



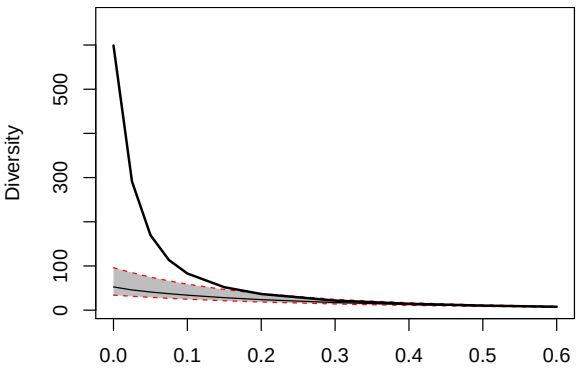

**(e)** Lognormal, 600 species, 500 individuals




**(f)** Geometric, 600 species, 500 individuals




**(g)** Lognormal, 600 species, 200 individuals




**(h)** Geometric, 600 species, 200 individuals

**Figure 19.** Estimation by the jackknife-unveiled estimator of the diversity profiles of simulated lognormal (left) and geometric (right) communities. The sample size decreases from 5000 (top) to 200 (bottom) individuals. The 95% confidence envelope of the estimation is shaded. The real diversity is plotted by the bold line.

## Appendix 3: R code

The necessary code to reproduce all results of the paper is given here.

```
library("entropart")
library("lmodel2")
library("vegan")

############### Simulation parameters ###############
Distributions <- c("lnorm", "geom")
Richness <- c(300, 600)
SampleSizes <- c(200, 500, 1000, 5000)
Corrections <- c("UnveilJ", "UnveiliC", "UnveilC", "ChaoWangJost",
        "GenCov", "ChaoShen", "Grassberger", "ZhangGrabchak", "None")
q.seq <- c(seq(0, .1, 0.025), .15, seq(.2, .6, 0.1))
NumberOfSimulations <- 1000
Alpha <- 0.05
#####################################################


############### Generate Communities ################
# Log normal: sd similar to BCI
sdlog <- 2
# Geometric: each species takes 10% of the remaining stick
prob <- .1

# rDistribution generates a probability distribution
rDistribution <- function (Distribution, S)
{
  # Generate the distribution
  Ps <- switch(Distribution,
    lnorm = (rlnorm(S, 0, sdlog) -> Ns)/sum(Ns),
    geom = prob/(1-(1-prob)^S)*(1-prob)^(0:(S-1))
  )
  return(as.ProbaVector(Ps))
}

# Generate distributions
for (Distribution in Distributions) {
  for (S in Richness) {
    assign(paste("P", Distribution, S, sep=""), rDistribution(Distribution, S))
  }
}

# Plot distributions
for (Distribution in Distributions) {
  for (S in Richness) {
    # Figure - Whittaker plot
    plot.SpeciesDistribution(eval(as.name(paste("P", Distribution, S, sep=""))),
      Distribution=Distribution, ylab="Probablity")
  }
}

# Figure: Distributions lognormal and geometric
plot.SpeciesDistribution(Plnorm300, Distribution="lnorm", ylab="Probablity",
```

```
  ylim=c(1e-15, 1e-1), cex=.5)
points(Pgeom300, cex=.5)
lNs <- log(Pgeom300)
Rank <- 1:length(Pgeom300)
reg <- lm(lNs~Rank)
lines(Rank, exp(reg$coefficients[1]+reg$coefficients[2]*Rank), col = "red")
#######################################################



############## Sample coverage ####################
# Function to simulate communities and compare observed to real coverage
SCfig <- function(Ps, Size, NumberOfSimulations, Distribution, S)
{
  # Simulate communities of the chosen size according to the chosen probability distribution
  MCSim <- rCommunity(NumberOfSimulations, size=Size, NorP=Ps)
  # Sum the actual probabilities of observed species in each simulation
  RealC <- colSums(Ps * (MCSim$Nsi>0))
  plot(RealC, MCSim$SampleCoverage.communities, xlab="Real Sample Coverage",
    ylab="Estimated Sample Coverage")
  abline(a=0, b=1)
  abline(h=mean(MCSim$SampleCoverage.communities), lty=2)
  abline(v=mean(RealC), lty=2)
  return(MCSim)
}

# Test all distributions
for (Distribution in Distributions) {
  for (S in Richness) {
    Ps <- eval(as.name(paste("P", Distribution, S, sep="")))
    # Test all sizes
    for (Size in SampleSizes) {
      SCfig(Ps, Size, NumberOfSimulations, Distribution, S)
    }

    # Overall evaluation: 2 simulations for each community size between 200 and 5000
    RealC <- EstimC  <- numeric()
    for (Size in min(SampleSizes):max(SampleSizes)) {
      MCSim <- rCommunity(2, size=Size, NorP=Ps)
      RealC <- c(RealC, colSums(Ps * (MCSim$Nsi>0)))
      EstimC <- c(EstimC, MCSim$SampleCoverage.communities)
    }
    plot(RealC, EstimC, xlab="Real Sample Coverage", ylab="Estimated Sample Coverage")
      abline(a=0, b=1)
    # Type II regression between real and observed C
    print(paste(Distribution, "Distribution", S, "species"))
    print(reg <- lmodel2(RealC~EstimC, nperm=1000))
  }
}
#######################################################



############## Entropy and Diversity ###############
# Function to build the confidence envelope of the estimation
```

```
CommunityProfileBySize <- function(FUN, Ps, q.seq, Size,
                           NumberOfSimulations, Correction, Alpha)
{
  # Create a MetaCommunity made of simulated communities
  MCSim <- rCommunity(NumberOfSimulations, size=Size, NorP=Ps, CheckArguments = FALSE)
  ProgressBar <- txtProgressBar(min=0, max=NumberOfSimulations)
  Sims <- matrix(nrow=NumberOfSimulations, ncol=length(q.seq))
  # Loops are required for the progress bar, instead of:
  # Sims <- apply(MCSim$Nsi, 2, function(Nsi) CommunityProfile(FUN, Nsi, q.seq, ...)$y)
  for (i in 1:NumberOfSimulations) {
    Sims[i, ] <- sapply(q.seq, function(q) FUN(MCSim$Nsi[, i], q,
      Correction=Correction, CheckArguments = FALSE))
    setTxtProgressBar(ProgressBar, i)
  }
  Means <- apply(Sims, 2, mean)
  Vars <- apply(Sims, 2, var)

  # Quantiles of simulations for each q
  EstEnvelope <- apply(Sims, 2, quantile, probs = c(Alpha/2, 1-Alpha/2))
  colnames(EstEnvelope) <- q.seq
  Profile <- list(x = q.seq,
    y = Means,
    low = EstEnvelope[1,],
    high = EstEnvelope[2,],
    var = Vars,
    Coverage = MCSim$SampleCoverage.communities
  )
  class(Profile) <- "CommunityProfile"
  return (Profile)
}


# Calculate diversity profiles
for (Distribution in Distributions) {
  for (S in Richness) {
    Ps <- eval(as.name(paste("P", Distribution, S, sep="")))
    # Real Profile
    Values <- sapply(q.seq, function(q) Tsallis(Ps, q, CheckArguments = FALSE))
    # Graphical parameters
    yLim <- c(0.9*min(Values), 1.1*max(Values))

    for (Size in SampleSizes) {
      for (Correction in Corrections) {
        # Entropy profile
        CommunityProfileBySize(bcTsallis, Ps, q.seq, Size,
          NumberOfSimulations, Correction, Alpha) -> Env
        # Transform entropy into diversity
        DEnv <- expq.CommunityProfile(Env)
        DValues <-  sapply(1:length(q.seq), function(i) expq(Values[i], q.seq[i]))
        # Save the data
        Env$Real <- Values
        # Calculate RMSE
        Env$RMSE <- sqrt((Env$Real-Env$y)^2 + Env$var)/Env$Real
        assign(paste("E", Distribution, S, "_", Size, Correction, sep=""), Env)
      }
```

```
    }
  }
}

# Plot diversity profiles
for (Distribution in Distributions) {
  for (S in Richness) {
    for (Size in SampleSizes) {
      for (Correction in Corrections) {
        DEnv <- eval(as.name(paste("E", Distribution, S, "_", Size, Correction, sep="")))
        plot(DEnv, ylim=c(min(DEnv$Real)*.9, max(DEnv$Real)*1.1), LineWidth=1, main="")
        lines(q.seq, DEnv$Real, lwd=2)
      }
    }
  }
}
#####################################################



###### Figure DP lnormal300 - 1000 individuals #####
plot(Elnorm300_1000ChaoWangJost$x, Elnorm300_1000ChaoWangJost$y,
  ylim=c(min(Elnorm300_1000ChaoWangJost$Real)*.9,
  max(Elnorm300_1000ChaoWangJost$Real)*1.1),
  main="", xlab = "Order of Diversity", ylab = "Diversity", xlim=c(0, 0.6), type="n")
CEnvelope(Elnorm300_1000ChaoWangJost,
  LineWidth=2, main="", xlim=c(0, 0.6), ShadeColor=NA, col="red")
lines(Elnorm300_1000ChaoWangJost$x, Elnorm300_1000ChaoWangJost$high,
  col="red", lty=1)
lines(Elnorm300_1000ChaoWangJost$x, Elnorm300_1000ChaoWangJost$low,
  col="red", lty=1)
CEnvelope(Elnorm300_1000UnveilJ, lty=2, LineWidth=2, col="blue",
  BorderColor="blue", ShadeColor=NA)
lines(q.seq, Elnorm300_1000ChaoWangJost$Real, lwd=2, lty=1)
#####################################################



####### Figures RMSE ################################
# Figure RMSE lnorm300
plot(q.seq, seq(0, 1, length.out=length(q.seq)), type="n",
  xlab="Order of Diversity", ylab="RMSE", xlim=c(0, 0.6))
ltype <- 1
Cases <- vector()
for (Correction in Corrections) {
  lines(q.seq,
    eval(as.name(paste("Elnorm300_1000", Correction, sep="")))$RMSE, lty=ltype)
  ltype <- ltype+1
  Cases <- append(Cases, Correction)
}
legend("topright", legend = Cases, lty = 1:ltype)


# Figure RMSE geom300
plot(q.seq, seq(0, 1, length.out=length(q.seq)), type="n",
```

```
  xlab="Order of Diversity", ylab="RMSE", xlim=c(0, 0.6))
ltype <- 1
Cases <- vector()
for (Correction in Corrections) {
  lines(q.seq, eval(as.name(paste("Egeom300_1000", Correction, sep="")))$RMSE, lty=ltype)
  ltype <- ltype+1
  Cases <- append(Cases, Correction)
}
legend("topright", legend = Cases, lty = 1:ltype)
#####################################################



####### Real data ###################################
# BCI
data(BCI)
NsBCI <- as.AbdVector(colSums(BCI))
plot(NsBCI, Distribution="lnorm")
NBCI <- sum(NsBCI)
AbdBCI <- AbdFreqCount(NsBCI)
# Sample coverage
(CBCI <- Coverage(NsBCI))
S1BCI <- AbdBCI[which(AbdBCI[, 1] == 1), 2]
S2BCI <- AbdBCI[which(AbdBCI[, 1] == 2), 2]
# Confidence interval of the sample coverage estimation
ICBCI <- qt(1-Alpha/2, NBCI)*sqrt(S1BCI*(1-S1BCI/NBCI)+2*S2BCI)/NBCI
# Number of species
Richness(NsBCI, Correction="Chao1")
Richness(NsBCI, Correction="Jackknife")
# Diversity profile
DPBCI <- CommunityProfile(Diversity, NsBCI,
  NumberOfSimulations = 100, Alpha=Alpha)
plot(DPBCI)

# Paracou 18
data(Paracou618)
NsP18 <- as.AbdVector(Paracou618.MC$Nsi[, 2])
plot(NsP18, Distribution="lnorm")
NP18 <- sum(NsP18)
AbdP18 <- AbdFreqCount(NsP18)
# Sample coverage
(CP18 <- Coverage(NsP18))
S1P18 <- AbdP18[which(AbdP18[, 1] == 1), 2]
S2P18 <- AbdP18[which(AbdP18[, 1] == 2), 2]
# Confidence interval of the sample coverage estimation
ICP18 <- qt(1-Alpha/2, NP18)*sqrt(S1P18*(1-S1P18/NP18)+2*S2P18)/NP18
# Number of species
Richness(NsP18, Correction="Chao1")
Richness(NsP18, Correction="Jackknife")
# Diversity profile. Very variable, 1000 simulations required.
DPP18 <- CommunityProfile(Diversity, NsP18, Correction="UnveilJ",
  NumberOfSimulations = 1000, Alpha=Alpha)
plot(DPP18)
```